A focused information criterion for graphical models in fMRI connectivity with high-dimensional data

Eugen Pircalabelu[1], **Gerda Claeskens**[1], **Sara Jahfari**[2] and **Lourens J. Waldorp**[3]

[1]KU Leuven, ORSTAT and Leuven Statistics Research Center, Faculty of Economics and Business, Naamsestraat 69, 3000 Leuven, Belgium
eugen.pircalabelu@kuleuven.be; gerda.claeskens@kuleuven.be

[2] Vrije Universiteit Amsterdam, Department of Cognitive Psychology, Faculty of Psychology and Education, Van de Boechorststraat 1, 1081 BT Amsterdam, The Netherlands
s.jahfari@vu.nl

[3]University of Amsterdam, Department of Psychological Methods, Faculty of Social and Behavioural Sciences, Weesperplein 4, 1018 Amsterdam, The Netherlands
waldorp@uva.nl

Abstract

Connectivity in the brain is the most promising approach to explain human behavior. Here we develop a focused information criterion for graphical models to determine brain connectivity tailored to specific research questions. All efforts are concentrated on high-dimensional settings where the number of nodes in the graph is larger than the number of samples. The graphical models may include autoregressive times series components, they can relate graphs from different subjects, or pool data via random effects. The proposed method selects a graph with a small estimated mean squared error for a user-specified focus. The performance of the proposed method is assessed on simulated datasets and on a resting state functional magnetic resonance imaging (fMRI) dataset where often the number of nodes in the estimated graph is equal to, or larger than the number of samples.

Keywords: fMRI connectivity; Focused information criterion; Model selection; Gaussian graphical model; Penalization; High-dimensional data.

1 Introduction

Connectivity based on measurements with functional magnetic resonance imaging (fMRI), is thought to be one of the key methods to clarify and understand how regions within our brain relate and communicate. Several methods have been used to determine connectivity from fMRI resting state data (where subjects do not perform any task), such as pairwise (Pearson) correlations, partial correlations (O'Neil et al., 2014), dynamic causal modeling (Friston et al., 2014), structural equation modeling (James et al., 2009), and Granger causality (Deshpande et al., 2011). Here we propose an approach based on graphical models for the analysis of resting state fMRI data, where one based on theory (i.e., genetic profile) or empirical data can define an *a priori* focus or emphasis on a sub-selection of regions.

Because resting state connectivity analysis is often performed across the whole brain, with a large number of brain regions (nodes), many of these studies require a large number of parameters (connections between nodes) while having only a limited number of observations (recorded time points), the so-called high-dimensional or $p > n$ setting. While this purely exploratory fashion of connectivity analysis remains of importance, relationships between resting state connectivity patterns and more specific cognitive abilities (or malfunctions) might benefit from a more focused approach, with an emphasis on a specific set of brain regions. For example, executive control – i.e., our ability to voluntary or strategically control or select planned actions – has been repeatedly linked (among others) to regions within the prefrontal cortex in both primates (e.g., Isoda and Hikosaka, 2007) and humans (e.g., Frank, 2011; Ridderinkhof et al., 2004; Jahfari et al., 2011, 2012). It is then desirable to inform the analysis somehow of the preference for the regions in the prefrontal cortex and the focused information criterion offers a solution to this issue.

To select a graphical model we use a series of repeated regressions of all nodes involved (neighborhood selection, see Meinshausen and Bühlmann, 2006) and then select a model according to an information criterion. When the number of observations is less than the number of parameters to estimate, a penalty is required to obtain estimates of the parameters. Popular choices are the graphical lasso (Friedman et al., 2008) and its variants such as the adaptive lasso (Fan et al., 2009), or the shrinkage estimator introduced by James and Stein (1961). Here we combine the choice of the penalty with model selection. A model is selected using an extension of the focused information criterion (FIC) by Claeskens and Hjort (2003), which minimizes the estimated mean squared error of an estimator of a particular function of the parameters, the focus. The focus can be different for different research questions. Pircalabelu et al. (2015) use the FIC to estimate graphical models with a small number of nodes. The goal of the present paper is to augment and extend the use of FIC to high-dimensional graphical models with several popular penalties. We evaluate the overall differences and similarities between these penalties, and select different regularization levels in a data-driven way by minimizing MSE expressions. Using the focused information criterion for model selection of brain connectivity does not give guarantees with respect to the true underlying graph. A model selected by the FIC does not necessarily contain all true edges. See Bühlmann (2013) for a discussion concerning linear models and screening properties. Neither is such a model necessarily consistent. The benefits of the FIC are that (i) the mean squared error of the estimator for a particular function of the parameters of interest (focus) is minimized, (ii) by doing so the prediction error related to that focus is minimized, and (iii) the idea of a 'single model fits all' is relaxed. The usual approach when dealing with settings where $p > n$, is to combine the estimation with model selection by using sparsity enforcing penalties, which have the direct objective of setting parameters to zero, ensuring thus a sparse solution. The most popular such penalty is the ℓ_1 penalty which in the context of estimating graphs is used to make the decision if an edge should or should not be present in the estimated graph. We propose to separate the estimation from the model selection, as follows. Forced by the high-dimensional context where $p > n$, a penalization method is required in order to estimate the parameters. To decide if an edge should be present in the estimated graph, we rely on the model selection mechanism associated with the FIC, as opposed to relying on the sparsity properties of the penalty. We estimate the final graph by scoring various configurations of edges using the FIC value and we keep modifying the graph until the FIC value is optimized.

Due to the importance of the prefrontal cortex (PFC) in goal-oriented behavior and executive control tasks (Ridderinkhof et al., 2004), resting state studies interested in executive control functions might benefit from a specific focus on PFC regions. One of the focuses used in the data analysis is, therefore, one where regions in the prefrontal cortex are emphasized. This entails that the estimated edges in that specific part of the brain should have lower mean squared error than edges between other regions, and as such be more accurate. The FIC does exactly this. The choice for minimization of the mean squared error is justified since it provides a good way to balance squared bias and variance, in other words, fit and generalization, respectively.

Additionally, by using as focus the observed measurements at a certain time point k, we estimate with the FIC a network that is designed to perform well with respect to the MSE of this focus at that time point. By varying the time point it is possible that a different network is obtained since the focus has changed. By repeating the process for different time points we can inspect possible changes over time.

In Abegaz and Wit (2013) vector autoregressive time series are used as models for time-varying networks, where lag-1 time points are incorporated. Such autoregressive effects are also included in our models. Zhou et al. (2010) also assume autoregressive processes underlying the changes over time in networks. (Kolar et al., 2010), in contrast, have a model that allows abrupt changes in time with the restriction that the total variation is bounded.

In the brain imaging community, the time varying and dynamical functional connectivity has been investigated in the works of Allen et al. (2014), Cribben et al. (2012) and Leonardi et al.

(2013), among others.

The 'default mode network' (DMN) is often studied in resting state fMRI. Many cognitive states in psychology have been linked to the DMN regions (see Raichle et al., 2001). The regions forming the DMN did not emerge on the basis of choice; i.e. they were not chosen *a priori*, but emerged from research as a set of regions found active when participants were not involved in a task. (Honey et al., 2009) have shown that in the default mode network there exists a connection between functional and structural connectivity. In the literature (for a review see Buckner et al., 2008) the DMN has been linked to attention disorders, monitoring the external environment, self-reflective thought and judgment, autism, schizophrenia and Alzheimer's disease. Also the fronto-occipital (FO) connections are thought to be especially important for schizophrenia (Bassett et al., 2008; Chai et al., 2011; Woodward et al., 2011).

Many issues are involved in determining the regions of interest (ROI) from fMRI data (see e.g., Lindquist, 2008; Waldorp, 2009). One of the issues is that contiguous voxels in a volume are spatially related. This issue can be taken up in several ways, all involving a model for the spatial distribution of brain activity (e.g., Weeda et al., 2010). Here we take the common approach of using atlas based ROIs that have been aggregated over different subjects (see e.g., Hagmann et al., 2008; Honey et al., 2009). We refer the reader to Appendix B for the names of the regions used in the study. The dataset contains information on partitions of these regions.

More information about the data and the acquisition procedure is offered in Section 4.

The rest of the paper is organized as follows. The general methodology of FIC with some background information is presented in Section 2, followed by a simulation study in Section 3. Section 4 contains the obtained results from the analysis on an fMRI dataset. Some extensions are presented in Section 5 and Section 6 concludes.

2 The proposed FIC method

Consider a p-dimensional multivariate random variable $X = (X_1, \ldots, X_p)$, which is normally distributed with a certain mean vector and covariance matrix Σ. Assuming a non-singular matrix Σ, there is a one-to-one mapping between the conditional independencies that hold in the distribution and a graphical structure $\mathcal{G}(\mathcal{E}, \mathcal{V})$, with nodes in $\mathcal{V}$ and edges in $\mathcal{E}$. Each of the univariate variables $X_1, \ldots, X_p$ corresponds to one node in the set $\mathcal{V}$ and the set of edges $\mathcal{E}$ is a subset of pairs of distinct nodes in $\mathcal{V} \times \mathcal{V}$. Lauritzen (1996) showed that if X_i is independent of X_j conditionally on all remaining variables in the model, denoted by $X_i \perp X_j | X_{\{1,\ldots,p\}\setminus\{i,j\}}$, then the pairs $(i,j) \cup (j,i) \notin \mathcal{E}$. Independence in the Gaussian case implies that $\Sigma^{-1}_{ij} = \Sigma^{-1}_{ji} = 0$. In other words, edges in $\mathcal{E}$ can be obtained by estimating the non-zero elements in the inverse covariance matrix, also called concentration matrix, a property known in the literature as 'covariance selection' (Dempster, 1972).

An estimate of the inverse covariance matrix can be obtained in several ways. First, the graphical Lasso (GL) maximizes the penalized log-likelihood of the data using as penalty $\lambda||\Sigma^{-1}||_1$, where the ℓ_1 norm of a matrix is the sum of the absolute values of the matrix entries (see, e.g., Friedman et al., 2008; Witten et al., 2011; Mazumder and Hastie, 2012; Yuan and Lin, 2007; Ravikumar et al., 2008; Krishnamurthy et al., 2012; Banerjee et al., 2008). Depending on the value of λ, the ℓ_1 penalized problem forces some elements in the concentration matrix to be set to 0, thus ensuring some degree of sparsity. An alternative is to take the ℓ_1 norm of the elementwise product $T \odot \Sigma^{-1}$ (Scheinberg and Rish, 2010; Li and Toh, 2010), which offers more flexibility. An extension to including lag-1 for time series data providing directed edges was proposed by Abegaz and Wit (2013). Following Dahlhaus and Eichler (2003) and (Gao and Tian, 2010) those authors then proceed at constructing a graph $\mathcal{G}$ by representing non-zero elements of the concentration matrix as undirected edges and non-zero autoregressive coefficients as directed edges. We adapt this approach for focused graph selection for the fMRI data, see Section 4.

Second, 'neighborhood selection' became popular with the work of (Meinshausen and Bühlmann, 2006). This procedure analyzes each node i sequentially, and estimates its neighborhood (ne_i), namely the smallest subset of nodes which conditioned upon, makes the current node independent of all remaining nodes. Once all neighborhoods are estimated, an estimated edge set is obtained using the 'AND' rule, or the 'OR' rule (Meinshausen and Bühlmann, 2006; Wainwright et al., 2007; Schmidt et al., 2007).

The application of penalized estimation methods in fMRI studies has been proposed in many other works. Examples of such applications include (Ryali et al., 2010, 2012), (Bunea et al., 2011) and (Lei et al., 2013), to cite just a few more recent applications.

Here we propose to estimate a graph that is optimal in the mean squared error (MSE) sense. We use the framework of neighborhood selection, where in each regression model misspecification is allowed (Claeskens and Hjort, 2003), and the likelihood function, including a penalty, is used to estimate the model parameters. Then the focus, emphasizing particular regions or pathways of the network, is used to determine the score of the focused information criterion (FIC) to determine which of the parameters are nonzero, in line with (Zhang and Liang, 2011) and (Claeskens, 2012). To determine the network we combine the FIC scores of all regressions for each of the nodes which make up the focus' estimate of MSE.

2.1 Likelihood and model specification

Consider a dataset consisting of n independent cases for each variable in the vector X. In a neighborhood selection framework, let in turn X_j be the response variable Y, and denote all remaining variables $\{X_i; i \in \mathcal{V} \setminus j\}$ by $\tilde{X}$. We denote the observed values for all remaining nodes for a case k where $k = 1, \ldots, n$ as $\tilde{x}_k$. The vector $\tilde{x}_k$ is further subdivided in two vectors: w_k for covariates that are always in the model and z_k for covariates that are subject to variable selection. Likewise, the parameters in the model are denoted as (θ, γ), corresponding to the vectors w_k and z_k, respectively.

The θ components correspond to the protected nodes, whose observed values are denoted by the vector w. The γ components correspond to the unprotected nodes whose observed values are denoted by the vector z. Considering nodes as protected or unprotected is entirely a researcher's decision, in the sense that one decides beforehand, informed by theory and research objectives, which nodes should always be included in the final model, i.e., the protected nodes, and from which sets of nodes the algorithm is allowed to select plausible ones, i.e., the unprotected nodes. If one knows that node i should be a neighbor of node j, then one would include it in the protected set, rather that letting the procedure decide if it should be included or not. In the narrow model, containing only protected variables, γ_0 is set to 0; in the full model all variables are included.

We assume the local misspecification framework, which assumes working with the density $f(y_k|w_k, z_k, \theta_0, \gamma_0 + \delta/\sqrt{n})$, where f is two times continuously differentiable in a neighborhood of the vector (θ_0, γ_0). The true unknown parameter vector $(\theta_0, \gamma_0 + \delta/\sqrt{n})$ is a vector of length $d = d_\theta + d_\gamma$, where d_θ and d_γ represent the lengths of the corresponding vectors θ and γ. The vector δ controls the size of the 'neighborhood' around the narrow model. In the basic model f is the normal density with $Y_k \sim N(w_k^\mathsf{T}\theta + z_k^\mathsf{T}(\gamma_0 + \delta/\sqrt{n}), \sigma^2)$, $k = 1, \ldots, n$ independent random variables with different means and a common variance. When pooling data from several subjects, we allow for extensions were $Y_1, \ldots, Y_n$ are correlated.

We define a *focus parameter* as a predetermined differentiable function $\mu(\theta, \gamma)$ which depends directly on the parameters of the density function and which is used to search for the estimator with the smallest MSE. The focus represents a mathematical translation of the research question, in the sense that the objective of the analysis is represented mathematically by the focus parameter. This is the quantity that we wish to estimate well, with small mean squared error. An estimator for this quantity is obtained by plugging-in the estimated values for θ and γ in the function μ. Under the

above local misspecification framework, $\mu_{\text{true}} = \mu(\theta_0, \gamma_0 + \delta/\sqrt{n})$. For example, the focus $\mu(\theta, \gamma) = E(Y_k|w_k, z_k)$ represents the expectation of a ROI at time k. Setting $E(Y_k|w_k, z_k) = w_k^\mathsf{T}\theta + z_k^\mathsf{T}\gamma$ then describes the expected value of that ROI at time k as a linear function. There are many possible choices as a focus parameter μ but it has to be a function of the parameters of the density and be differentiable.

For the high-dimensional setting in neighborhood selection, an estimator for (θ, γ) is obtained by maximizing the penalized objective function with respect to θ and γ,

$$Q(\theta, \gamma) = \frac{1}{n}\sum_{k=1}^{n} \log f(y_k|w_k, z_k, \theta, \gamma) - \frac{\lambda_n}{n}\sum_{j=1}^{d_\gamma} \psi(|\gamma_j - \gamma_{j0}|), \tag{2.1}$$

for a given penalty function ψ (that is twice differentiable in 0) and an external value λ_n. An estimator obtained with (2.1) is denoted by $(\hat{\theta}, \hat{\gamma})$. As an example, consider the BOLD (blood-oxygen-level dependent) responses from fMRI used for the analysis of connectivity which are correlated across time points (Worsley, 2001). A popular method to tackle temporal dependence is a Gaussian autoregressive AR(1) model. In our setting we require the model's intercept and error variance to be in all models

$$Q(\theta, \gamma) = \frac{-n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \sum_{k=2}^{n} \frac{y_k - \alpha - \tilde{x}_k^\mathsf{T}\beta - \rho y_{k-1}}{2\sigma^2} - \frac{\lambda_n}{n}\{\sum_{j=1}^{d_\gamma} \psi(|\beta_j - \beta_{j0}|) + \psi(|\rho - \rho_0|)\},$$

where $\theta = (\sigma^2, \alpha)$ and $\gamma = (\rho, \beta)$. More examples and extensions are given in Section 5.

2.2 FIC for penalized estimation of nodewise models

Since θ is present in all models, we concentrate the model selection process on γ and thus the penalty in (2.1) is applied only to γ. We will always restrict $d_\theta < n$, but allow that $d_\gamma > n$ (although it cannot grow with n).

Let $\hat{\mu} = \mu(\hat{\theta}, \hat{\gamma})$ be the penalized maximum likelihood estimator of the focus, obtained by evaluating μ at the estimated values $(\hat{\theta}, \hat{\gamma})$. For simplicity, we suppress in the notation the dependency of $\hat{\theta}$ and $\hat{\gamma}$ on the penalty λ_n. The objective is to estimate the focus $\mu(\theta, \gamma)$ in the 'best' way, here defined in terms of MSE. We proceed by estimating μ based on different models (i.e. different configurations of neighbors for the node under consideration) and denote the estimated quantity using model S as $\hat{\mu}_S = \mu(\hat{\theta}_S, \hat{\gamma}_S, \gamma_{0,S^c})$. Note that S is a subset of indices corresponding to all remaining nodes and S^c denotes the complementary set. The length of the vector $\hat{\theta}_S$ is always equal to d_θ, but the actual value of the estimator may depend on which of the components of the vector γ are included in the index set S. The vector $\hat{\gamma}_S$ estimates the components of γ that are included in S, the other components are set to zero, they form the vector γ_{0,S^c}. The cardinality of the set $S \cup S^c$ is thus d.

For each model indexed by S, based on the above quantities, we have (see Claeskens, 2012)

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \xrightarrow{\mathcal{D}} \Lambda_S \sim N(\text{Mean}(\mu, S, \delta, c), \text{Var}(\mu, S)). \tag{2.2}$$

The mean depends on the chosen focus μ, the submodel S, the value of δ indicating the distance between the parameters from the simplest model and the true model, and on the chosen penalization via $c = \lambda_0\psi''(0)1_q$, where $\lambda_n/\sqrt{n} \to \lambda_0 > 0$. The variance depends only on the focus μ and on the submodel S. Precise values of the mean and variance are defined in Appendix A. By assumption μ is a differentiable function for which the partial derivatives with respect to θ and γ exist.

As is well-known, the MSE can be decomposed into the squared bias and variance, which in the case of Λ_S in (2.2) gives

$$\text{MSE}(\hat{\mu}_S) = \text{Mean}(\mu, S, \delta, c)^2 + \text{Var}(\mu, S). \tag{2.3}$$

Appendix A contains the exact MSE expression based on (2.2) which is used for the implementation. To estimate $\text{MSE}(\hat{\mu}_S)$ in (2.3) we proceed by plugging-in the empirical version of the unknown quantities using parameter estimates from the full model. An estimator for δ is $\hat{\delta} = \sqrt{n}(\hat{\gamma}_{\text{full}} - \gamma_0) \to_{\mathcal{D}} N(\delta, J^{11})$, with J^{11} a submatrix of the inverse of the Fisher information matrix, see the appendix for a precise definition. Since the squared bias is needed, an unbiased estimator for $\delta\delta^\mathsf{T}$ is $\hat{\delta}\hat{\delta}^\mathsf{T} - \hat{J}^{11}$, since $E(\hat{\delta}\hat{\delta}^\mathsf{T}) = \delta\delta^\mathsf{T} + J^{11}$. Hence, to estimate the MSE we require $\hat{\delta}$ and $\hat{J}^{11}$.

Since the MSE is defined per node, we define the FIC for the entire estimated graph as the sum of MSEs, where each node $l \in \mathcal{V}$ has a particular model S_l based on which we have constructed the estimator $\hat{\mu}_{l;S_l}$. Denote the set $\mathcal{S} = \{S_1, \ldots, S_p | S_1 \subseteq \{\mathcal{V} \setminus 1\}; \ldots; S_p \subseteq \{\mathcal{V} \setminus p\}\}$, then

$$\text{FIC}(\mathcal{G}(\mathcal{E}_\mathcal{S}, \mathcal{V})) = \sum_{l=1}^{p} \widehat{\text{MSE}}(\hat{\mu}_{l;S_l}). \tag{2.4}$$

The objective is to minimize (2.4) over the set $\mathcal{S}$.

2.3 Steps to obtain an FIC graph based on nodewise models

Since we are dealing with fMRI time series, in each nodewise regression we incorporate both instantaneous and lag-1 effects in the network, resulting in a regression with $2p - 1$ predictors for each nodewise model. Once all nodewise models are selected, we apply the following 'OR' rule adapted from Meinshausen and Bühlmann (2006):

$$\begin{array}{lcll}
\hat{\mathcal{E}}^{\lambda,\text{OR}}_{i-j} & = & \{(i,j) \cup (j,i) : i_k \in \hat{ne}^\lambda_{j_k} \text{ OR } j_k \in \hat{ne}^\lambda_{i_k}\} & \text{instantaneous effects; undirected edges} \\
\hat{\mathcal{E}}^{\lambda,\text{OR}}_{i\to j} & = & \{(i,j) : i_{k-1} \in \hat{ne}^\lambda_{j_k}\} & \text{lag 1 effects; directed edges} \\
\hat{\mathcal{E}}^{\lambda,\text{OR}}_{i\leftarrow j} & = & \{(j,i) : j_{k-1} \in \hat{ne}^\lambda_{i_k}\} & \text{lag 1 effects; directed edges} \\
\hat{\mathcal{E}}^{\lambda,\text{OR}} & = & \left\{\hat{\mathcal{E}}^{\lambda,\text{OR}}_{i-j} \cup \hat{\mathcal{E}}^{\lambda,\text{OR}}_{i\to j} \cup \hat{\mathcal{E}}^{\lambda,\text{OR}}_{i\leftarrow j}\right\} & \text{combined directed and undirected edges,}
\end{array}$$

where $\hat{ne}^\lambda$ denotes the neighborhood of the considered node for a certain value of λ.

The main steps of our procedure are summarized as follows:

1. Specify the focus μ. Decide on the likelihood and penalty function to construct the penalized function $Q(\theta, \gamma)$.

2. At each node and for a set of explanatory models for that node, estimate the MSE of the focus estimator by FIC and choose the model that minimizes FIC. This requires optimizing $Q(\theta, \gamma)$ for different models and estimating the quantities needed to construct the FIC for this focus. In our approach we find that model in a greedy forward stepwise manner, where we start by evaluating all one-variable models, select the best performing one and then add one more variable at each step until the FIC value for that node cannot be improved.

3. The 'OR' rule is applied to construct a graph from the nodewise models. We add an undirected edge indicating a contemporaneous relation between two nodes if at least one node is part of the other node's selected model. A directed edge indicating a temporal, lagged relation is added between two nodes if the lag 1 effect of one node is part of the selected model for the other node. An FIC estimated mixed graph results.

These steps are applicable to the different choices of criterion functions $Q(\theta, \gamma)$ and allow for different likelihoods and for different penalty functions.

2.4 The choice of penalty function

In principle, any type of penalty such as an ℓ_1 (Meinshausen and Bühlmann, 2006) or bridge (Fu, 1998), elastic net (Zou and Hastie, 2005), adaptive lasso (Zou, 2006), SCAD (Fan and Li, 2001), etc. can be used, see below. Although we require the use of a penalty because we are dealing with the high-dimensional setting where $d_\gamma > n$ is possible, a *sparsity* enforcing penalty is unnecessary. The reason is that exclusion/inclusion of a certain predictor, which translates into a zero/non-zero γ component, is determined by the value of the FIC.

Since differentiability of ψ is needed, for non-differentiable functions we proceed as in Fan and Li (2001) and replace ψ by a local quadratic approximation (LQA), which has the advantage of being low in computational complexity, as an iterative Newton-Raphson algorithm can be employed for optimization purposes. To improve numerical stability one can also introduce small 'perturbations' as in Hunter and Li (2005). The local linear approximation (LLA) of Zou and Li (2008), not used in this paper, could be an alternative.

With LQA, $\psi(|\gamma_j - \gamma_{j0}|)$ is approximated by a Taylor expansion and its first and second partial derivatives (with respect to $\gamma_j - \gamma_{j0}$) are approximated by

$$\begin{aligned}
\psi(|\gamma_j - \gamma_{j0}|) &\approx \psi(\gamma_{j,\text{apx}}) + \tfrac{1}{2}\frac{\psi'(|\gamma_{j,\text{apx}}|)}{|\gamma_{j,\text{apx}}|}\left[(\gamma_j - \gamma_{j0})^2 - \gamma_{j,\text{apx}}^2\right];\\
\psi'(|\gamma_j - \gamma_{j0}|) &\approx \frac{\psi'(|\gamma_{j,\text{apx}}|)}{|\gamma_{j,\text{apx}}|}(\gamma_j - \gamma_{j0});\\
\psi''(|\gamma_j - \gamma_{j0}|) &\approx \frac{\psi'(|\gamma_{j,\text{apx}}|)}{|\gamma_{j,\text{apx}}|},
\end{aligned}$$

for $\gamma_{j,\text{apx}}$ an approximation point close to $(\gamma_j - \gamma_{j0})$. With this approximation, several penalties can be used in (2.1), including

- lasso: $\psi_l(|\gamma_j - \gamma_{j0}|) = |\gamma_j - \gamma_{j0}|$;
- bridge: $\psi_b(|\gamma_j - \gamma_{j0}|) = |\gamma_j - \gamma_{j0}|^\alpha;\ \alpha > 0$;
- hard thresholding: $\psi_h(|\gamma_j - \gamma_{j0}|) = \lambda^2 - (|\gamma_j - \gamma_{j0}| - \lambda)^2 I(|\gamma_j - \gamma_{j0}| < \lambda)$;
- adaptive lasso: $\psi_{al}(|\gamma_j - \gamma_{j0}|) = w_j|\gamma_j - \gamma_{j0}|$, for a weight w_j;
- SCAD (first derivative):
  $\psi_s'(|\gamma_j - \gamma_{j0}|) = I(|\gamma_j - \gamma_{j0}| \le \lambda) + \frac{(a\lambda - |\gamma_j - \gamma_{j0}|)_+}{(a-1)\lambda} I(|\gamma_j - \gamma_{j0}| > \lambda); a > 2.$

The ℓ_2 penalty has the convenient advantage that a closed form estimator exists, it is differentiable, and leads to tractable mean squared error expressions for the focus estimators making the bias-variance trade-off explicit.

2.5 Regularization level λ

Given one of the above penalties and a corresponding value of $\psi''(0)$, we propose to choose the regularization parameter λ by solving a mean squared error minimization problem. In particular the regularization parameter that we propose to use is the one that minimizes $\text{MSE}(\hat{\mu}_S)$ in (2.3). Since this is a quadratic function in $c = \lambda_0\psi''(0)1_q$, we solve for c in the equation $\partial\text{MSE}(\hat{\mu}_S)/\partial c = 0$. For $\psi''(0) \neq 0$, the optimal regularization level is obtained as $\hat{\lambda}_S = \arg\min_c \text{MSE}(\hat{\mu}_S)\sqrt{n}/\psi''(0)$ which leads to an explicit expression of a model dependent value $\hat{\lambda}_S$, given in (A.2).

Since the $\hat{\lambda}_S$ depends on δ, appearing in the MSE, we are faced with an endogeneity problem: to use the optimal $\hat{\lambda}_S$ we need to know $(\hat{\theta}, \hat{\gamma})$, but in order to estimate the two unknown vectors we need λ_S. One solution to the problem is the following two-step procedure: we first estimate $(\hat{\theta}, \hat{\gamma})$ on a grid of λ values, and then select the optimal estimates based on the GCV criterion (see Craven and Wahba, 1978). We mention that using this value for λ we estimate the quantities $\hat{\theta}, \hat{\gamma}$

Algorithm 1 Nodewise MSE calculations

1. Specify the focus of interest for a node l in the graph, represented generically by the variable Y. For example, $\mu(\theta,\gamma) = E(Y_k|w_k, z_k)$ where (θ,γ) represents the parameters of the underlying density and (w_k, z_k) represents the measurements for all other nodes at a fixed datapoint (either in-sample or out-of-sample);
2. Choose the approximated penalty function ψ as in Section 2.4 and specify a value for λ;
3. Optimize (2.1) where autoregressive effects of an arbitrary order are allowed and obtain $(\hat{\theta},\hat{\gamma})$ in the full (most complex) model, after which construct $\hat{\delta} = \sqrt{n}\hat{\gamma}$;
4. Construct the ψ'' and evaluate it at 0.
5. Estimate the empirical Fisher information matrix $(\hat{J})$ and its inverse $(\hat{J}^{-1})$ at the full model;
6. Specify at the node l a collection of models, represented by the potential neighbor variables and their lagged versions (this is constructed incrementally in a forward manner in Algorithm 2);
7. For each model S (a configuration of potential neighbors and lagged counterparts) in the above collection, construct the empirical Fisher information matrix corresponding to model S $(\hat{J}_S)$ using the projection matrices π_S;
8. Compute the quantities $\hat{\omega} = \hat{J}_{10}\hat{J}_{00}^{-1}\frac{\partial\mu}{\partial\theta} - \frac{\partial\mu}{\partial\gamma}$, $\hat{G}_S = \hat{J}^{11,S,0}(\hat{J}^{11})^{-1}$, $\hat{J}^{11,S,0} = \pi_S^{\mathsf{T}}\hat{J}^{11,S}\pi_S$ based on the partitions of $\hat{J}$, $\hat{J}^{-1}$ and $\hat{J}_S$, the projection matrix π_S and the partial derivatives of the focus;
9. Compute λ_S as in (A.2) and $\text{MSE}(\hat{\mu}_S)$ as in (A.1) since all necessary quantities have been estimated in the previous steps.

which are necessary in the calculation of ω and $\hat{\delta}$. We then estimate $\hat{J}$, partition it according to the dimensions of the vectors θ and γ after which the matrix G_S is computed. All the necessary quantities for computing λ_S from (A.2) are now available. We proceed afterward by estimating $\hat{c}_S$ and the $\text{MSE}(\hat{\mu}_S)$ expression is immediately available for each model S where unknown quantities are estimated. See Algorithm 1 for more details on the implementation; all quantities mentioned are defined in Appendix A.1.

Another possibility to determine the regularisation level λ is by k-fold cross-validation, where $k = 10$, say. This would avoid using the likelihood to determine both the parameters and the regularisation level. However, as we show in the simulations, optimising the regularisation level through the likelihood, as described above, results in good performance.

2.6 An algorithmic view on estimating FIC graphs

To summarize the above procedures and the steps that are followed in estimating the FIC graphs, we provide in this section in an algorithmic format how one computes the estimated $\widehat{\text{MSE}}(\hat{\mu}_{l;S_l})$ values for a focus specified at the node l (see Algorithm 1), and how one searches for the configuration of instantaneous (undirected edges) and lagged (directed edges) effects (see Algorithm 2).

In Algorithm 1 one starts off by specifying the focus, the penalty function to be used and the regularization level λ. One then proceeds with estimating the Fisher information matrix and the parameters $\hat{\delta}$, $\hat{\omega}$ and $\hat{\gamma}$ after which all the necessary quantities for computing the MSE expression as in (A.1) can be directly computed. Plugging-in all the necessary quantities one obtains the estimated MSE expression for a model S at node l.

In Algorithm 2 we search in a forward manner for the nodewise model that optimizes the MSE expression. We start off by specifying an empty model for the node and compute the estimated MSE using Algorithm 1. We then modify the model and check if adding other nodes decreases the MSE values. We repeat the search until we have introduced in the optimal model S_l the nodes (or lagged versions of the nodes) that have resulted in the best MSE values. This approach is

Algorithm 2 Nodewise based FIC graph estimation

for $l \in$ `all nodes in` $\mathcal{G}$ **do**
  Current $\widehat{\text{MSE}}(\hat{\mu}_{l;S_l}) = \infty$
end for
$\hat{\mathcal{G}} \leftarrow$ `empty graph`
for $l \in$ `all nodes in` $\mathcal{G}$ **do**
  $S_l \leftarrow \emptyset$
  Current $\widehat{\text{MSE}}(\hat{\mu}_{l;S_l}) \leftarrow$ `compute` $\widehat{\text{MSE}}(\hat{\mu}_{l;S_l})$ `using Algorithm 1`
  $Flag \leftarrow False$;
  while $Flag = False$ **do**
    for $m \in$ `possible instantaneous and lag-1 neighbors of node` l **do**
      $S_l \leftarrow S_l \cup m$
      Neighbor$_m\widehat{\text{MSE}}(\hat{\mu}_{l;S_l}) \leftarrow$ `compute` $\widehat{\text{MSE}}(\hat{\mu}_{l;S_l})$ `using Algorithm 1`
    end for
    Optimal Neighbor$_m\widehat{\text{MSE}}(\hat{\mu}_{l;S_l}) \leftarrow$ minimum Neighbor$_m\widehat{\text{MSE}}(\hat{\mu}_{l;S_l})$
    if Optimal Neighbor$_m\widehat{\text{MSE}}(\hat{\mu}_{l;S_l}) <$ Current $\widehat{\text{MSE}}(\hat{\mu}_{l;S_l})$ **then**
      Current $\widehat{\text{MSE}}(\hat{\mu}_{l;S_l}) \leftarrow$ Optimal Neighbor$_k\widehat{\text{MSE}}(\hat{\mu}_{l;S_l})$
      $S_l \leftarrow S_l \cup m$
      `possible instantaneous and lag-1 neighbors of node` $l \leftarrow$ `{possible instantaneous and lag-1 neighbors of node` l`}` $\setminus m$
    else
      $Flag \leftarrow True$;
    end if
  end while
end for
for $(i, j) \in$ `all nodes in` $\mathcal{G}$ **do**
  $\hat{ne}^{\lambda}_{i_k} \leftarrow S_i \setminus$ `{lag-1 neighbors}` $\in S_i$
  $\hat{ne}^{\lambda}_{j_k} \leftarrow S_j \setminus$ `{lag-1 neighbors}` $\in S_j$
  $\hat{ne}^{\lambda}_{i_{k-1}} \leftarrow S_i \setminus$ `{instantaneous neighbors}` $\in S_i$
  $\hat{ne}^{\lambda}_{j_{k-1}} \leftarrow S_j \setminus$ `{instantaneous neighbors}` $\in S_j$
end for
`Construct` $\left\{\hat{\mathcal{E}}^{\lambda,\text{OR}}_{i-j}; \hat{\mathcal{E}}^{\lambda,\text{OR}}_{i\to j}; \hat{\mathcal{E}}^{\lambda,\text{OR}}_{i\leftarrow j}\right\}$
$\hat{\mathcal{E}}^{\lambda,\text{OR}} \leftarrow \left\{\hat{\mathcal{E}}^{\lambda,\text{OR}}_{i-j} \cup \hat{\mathcal{E}}^{\lambda,\text{OR}}_{i\to j} \cup \hat{\mathcal{E}}^{\lambda,\text{OR}}_{i\leftarrow j}\right\}$
$\hat{\mathcal{G}} \leftarrow$ `update` $\hat{\mathcal{G}}$ `based on` $\hat{\mathcal{E}}^{\lambda,\text{OR}}$

taken for each node in turn and this results in estimating for each node its set of neighbor nodes. With very large graphs the search technique might not scale efficiently. Once the sets have been estimated, the 'OR' rule is applied and a graph is then constructed using the optimal identified sets of neighbors. We mention that if one desires to estimate undirected graphs, one can use the same algorithms, but restrict the influence of the other nodes to only the instantaneous edges and disregard the autoregressive effects. Such a strategy would be useful for applications that do not include time dependencies.

3 Simulation study

To compare known methods of graph estimation and selection to our proposed FIC method, we performed a simulation study. First, we generated independent data from a multivariate normal

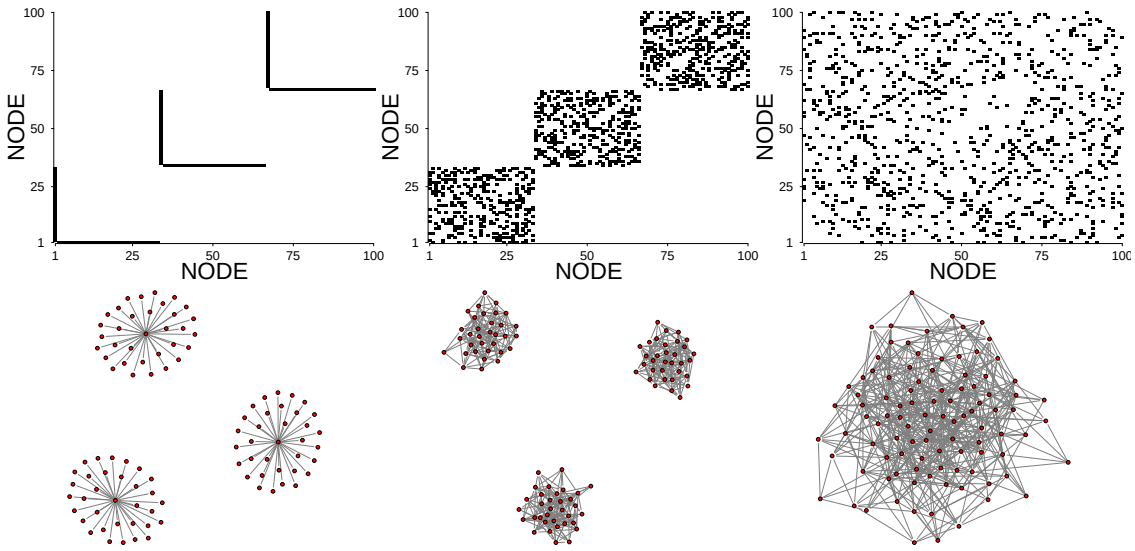


Figure 1: Simulated data. Visual representation of three Σ^{-1} matrices and their corresponding graph structure; hub in the left panel, cluster in the center panel, and random in the right panel. A black symbol (in the top row) signifies a non-zero entry in the Σ^{-1} on position (i, j) which is interpreted as a edge (in the bottom row) between two nodes. All blank entries are interpreted as non-existing edges between pairs of nodes.

distribution with three different undirected graphical structures: 'hub', 'cluster', and 'random' (see Figure 1). When hubs (left panel of Figure 1) or clusters (center panel of Figure 1) were used to generate the data, we have used 5 such structures and when a random graph structure (right panel of Figure 1) has been used, we have set the probability of connecting two nodes to 0.2. For each of the three graph structures, with $p = 35$ nodes we took settings with sample sizes $n = 15$, 30 and 75. With $p = 150$ nodes, we used sample sizes 75 and 300. For each scenario the number of simulation runs was set at 300. Out of these 15 different scenarios, there were 9 settings where the number of nodes was larger than the number of observations ($p > n$). For each scenario data have been generated either using constant variance at each node, set at 1 or using non-constant variances that were drawn at random from the interval [1, 2.5]. This leads to a total of 60 settings.

The competitive methods that have been studied here are FIC, GL, CLIME (Cai et al., 2011) and TIGER (Liu and Wang, 2012) as implemented in Li et al. (2013). For GL, CLIME and TIGER the regularization parameter has been chosen by three-fold cross-validation using the 'Likelihood' or 'Trace' loss

$$\text{Likelihood} = \text{trace}(\hat{\Sigma}^{-1}_{\text{train}}\hat{\Sigma}_{\text{test}}) - \log(|\hat{\Sigma}^{-1}_{\text{train}}|)$$
$$\text{Trace} = \text{trace}(\text{diag}(\hat{\Sigma}^{-1}_{\text{train}}\hat{\Sigma}_{\text{test}})^2 - I),$$

where $\hat{\Sigma}^{-1}_{\text{train}}$ is the concentration matrix estimated on the training sample and $\hat{\Sigma}_{\text{test}}$ is the covariance matrix fitted using the test set. The matrix $\text{diag}(\hat{\Sigma}^{-1}_{\text{train}}\hat{\Sigma}_{\text{test}})^2$ is the matrix $\hat{\Sigma}^{-1}_{\text{train}}\hat{\Sigma}_{\text{test}}$ whose diagonal elements are squared. The extended Bayesian information criterion (eBIC) developed in (Foygel and Drton, 2010), as well as the 'StARS' criterion implemented in Zhao et al. (2012) have been used for model selection.

For the FIC we defined the focus as $\mu(\theta, \gamma; \tilde{x}) = E[Y|\tilde{x}]$ where no variables are protected, and this is further specified to two focus choices for $\tilde{x}$:

$\mu_1 =$ $\mu(\theta, \gamma; \tilde{x})$ evaluated at the $\tilde{x}$ values corresponding to Huber's robust location of the center of the distribution;

$\mu_2 =$ $\mu(\theta, \gamma; \tilde{x})$ evaluated at the $\tilde{x}$ values that correspond to the median values of the measurements of each node.

Each of these two focuses has been treated once as an in-sample datapoint that was used in the training of the algorithms (in this case we have added the point to the original dataset) and once as an out-of-sample datapoint (the point was completely separated of the dataset used for training the algorithms).

For each of the two focuses, the FIC selects a graph and from the list of penalties described in Section 2.4 we have used the ℓ_2 penalty and the quadratic approximation to the ℓ_1 penalty to be more comparable with the competitor techniques. The regularization parameter has been chosen as described in Section 2.5.

For each method, once a graph is estimated, we estimate the elements of the corresponding concentration matrix, and construct the *empirical* MSE as

$$\overline{\text{MSE}} = \frac{1}{p} \sum_{l \in \mathcal{V}} \big(\tilde{x}_{0l} - \sum_{i \in ne(l)} \hat{\beta}_{li} X_i \big)^2,$$

where $\tilde{x}_0$ is the focus evaluation point and p is the number of nodes in the graph. The βs are estimated as follows: for GL, CLIME and TIGER based on the concentration matrix we construct at each node j a vector $(\Sigma^{-1}_{j,1}/\Sigma^{-1}_{j,j}, \ldots, \Sigma^{-1}_{j,m}/\Sigma^{-1}_{j,j})$ corresponding to the regression coefficients of all other m nodes in the regression model of node j (see Bühlmann and Van De Geer, 2011, pp. 436). Due to the sparsity nature of the techniques some components in this vector will be set to 0. For FIC we have used the penalized maximum likelihood estimator for (θ,γ) after optimizing (2.1).

We have used four measures to compare the performance of the methods: empirical MSE, sparsity, true positive rate (the number of correctly found edges divided by number of true edges) and false positive rate (the number of incorrectly found edges divided by number of true non-present edges). The sparsity of the estimated graphs has been estimated as $\text{SI} = 1 - |\hat{\mathcal{E}}|/(p(p-1)/2)$ where $|\hat{\mathcal{E}}|$ represents the number of estimated edges and p represents the number of nodes. A larger value corresponds to a 'sparser' graph which has a lower number of estimated edges.

3.1 Results

Pooled across all simulation runs from all settings the FIC-ℓ_2 and FIC LQA ℓ_1 estimated graphs produced the smallest empirical MSE values for each of the two focuses. From Table 2 we observe that for the first evaluation point the FIC provided better empirical MSE than the competitors regardless of it being an in-/out-of-sample point or whether the nodes had constant or non-constant variance. For the second evaluation point the FIC-ℓ_2 provided the best performance with the largest gains against the competitors for an in-sample evaluation point. For out-of-sample cases, its performance was closer to that of the competitors, but still better.

With respect to the sparsity of the estimated graphs, when compared to the competitor techniques, we observe that the FIC estimated graphs are not too sparse, but not too dense either. In general, since the FIC graph is estimated for a particular focus, it provided lower TPR rates when compared to the competitors that all estimate a global model that is not fine-tuned for the focus under analysis. In certain scenarios, the FIC can obtain better TPR rates than the average ones presented in Table 2. As an example, for the setting where data of size $n = 75$ were generated from a graph with cluster structure and $p = 35$ nodes, we observed an average TPR of 42%. Zooming-in on the sub-networks formed around a node (i.e. the node and all its neighbors) the TPRs in this setting ranged on average from 40% to 54%. The FPR rate was generally lower than for most of the competitor techniques. This suggests that the focus-tuned FIC graphs use only a part of the total true edges, the ones that reduce best the bias of the focus estimator without increasing too much the variance, such that in the end the MSE of the estimated graph is kept small.

	In;Ct		In;Non-ct		Out;Ct		Out;Non-ct	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
FIC - ℓ_2	0.002	0.014	0.005	0.028	0.002	0.028	0.005	0.062
FIC - LQA ℓ_1	0.003	0.016	0.006	0.029	0.003	0.042	0.008	0.101
Bayesian graph	0.057	0.010	0.100	0.030	0.056	0.011	0.100	0.035

Table 1: Simulated data. Empirical mean squared error (MSE) of estimated graphs for 2 focuses, pooled across 300 simulation runs and 36 different simulation settings for $p = 35$ (three types of graphs, 3 sample sizes). 'In/Out' refer to the settings where the point was an in-sample or out-of sample point while 'Ct/Non-Ct' refer to settings where the variance at each node was constant or non-constant (randomly sampled at each node).

Figure 2 presents the empirical MSE obtained as a function of λ. To investigate the change in the value of the empirical MSE of the estimated selection of the nodes with FIC when λ varies as compared to other methods we have simulated 300 datasets and took a sequence of ten values for λ ranging from 0.05 to 2.5. We have used here samples of size 75 for a random graph containing 35 nodes where the evaluation point was an out-of sample point for both cases of constant and non-constant variance at each node. Due to the two step procedure that we employ, the fluctuations in empirical MSE for this range of regularization values are milder for FIC than for the other two methods. The figure also suggests that increasing the penalty might be better for FIC, but we did not explore this any further since already for the larger values in the specified λ sequence the competitors provided empty graphs.

We have compared our method to a Bayesian graph learning technique with sparse priors, see Table 1. For this we generated 300 datasets using samples of size 15, 30 and 75 for random, hub and cluster graphs containing 35 nodes. The evaluation points were either in-sample or out-of sample points and the variance at each node was either constant for all nodes or non-constant. This leads to 36 settings. For fitting sparse undirected Bayesian graphs we have used the procedure implemented in Mohammadi and Wit (2015). Over all the simulation settings the FIC graphs were performing better than the Bayesian graphs for the first evaluation point. For the second evaluation point both techniques provided similar empirical MSE values when the point was an in-sample point, but the Bayesian graph was slightly better when the point was out-of-sample.

3.2 Dependent observations, lag-1

We have generated multivariate data from cluster, hub, and random structures though now with autoregressive effects of order 1. The probability of connecting two nodes was 0.2, the sample size was 30 or 75 and the number of nodes was 35. Autoregressive effects of order 1 (lag-1) have been used for modeling the mean structure as well as for generating the data. We compared FIC to the time series chain graphical model (TSCGM) from Abegaz and Wit (2013). The regularization parameter in TSCGM has been chosen using eBIC, the traditional BIC and a 'generalized' information criterion (GIC) defined as

$$\text{GIC} = -2\text{logLik} + \log(2p)\log\log(nm)[0.5\#\{\sigma_{ij}^{-1} > 0\} + p + \#\{\gamma_{ij} > 0\}],$$

where p is the number of nodes, n is the number of observations per time series, m is the number of time points and $\#\{\sigma_{ij}^{-1} > 0\}$, $\#\{\gamma_{ij} > 0\}$ are the number of non-zero entries in the matrices Σ^{-1} and Γ. The FIC has been used with an ℓ_2 penalty or a local quadratic approximation to the SCAD penalty (to be more comparable with the TSCGM methodology).

In order to obtain empirical MSE expressions, in the TSCGM we could use two strategies. Either we predict one node as a function of its past, through the directed structure and the matrix of AR(1) coefficients estimated by the method, or proceed by constructing the vectors of regression

	MSE								SI							
	In;Ct		In;Non-ct		Out;Ct		Out;Non-ct		In;Ct		In;Non-ct		Out;Ct		Out;Non-ct	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
FIC - ℓ_2	0.002	0.012	0.005	0.025	0.002	0.024	0.004	0.051	0.92	0.94	0.90	0.92	0.93	0.94	0.93	0.93
FIC - LQA ℓ_1	0.002	0.014	0.006	0.026	0.003	0.035	0.006	0.081	0.88	0.90	0.87	0.90	0.88	0.89	0.88	0.88
GL (CV-Likelihood)	0.020	0.035	0.054	0.082	0.018	0.032	0.049	0.075	0.69	0.69	0.85	0.85	0.71	0.71	0.87	0.87
GL (CV-Trace)	0.025	0.041	0.061	0.089	0.025	0.040	0.056	0.081	0.78	0.78	0.99	0.99	0.82	0.82	1.00	1.00
GL (StARS)	0.027	0.041	0.059	0.087	0.023	0.037	0.053	0.079	0.90	0.90	0.95	0.95	0.89	0.89	0.94	0.94
GL (eBIC)	0.033	0.049	0.062	0.090	0.032	0.047	0.056	0.082	0.96	0.96	1.00	1.00	0.96	0.96	1.00	1.00
CLIME (STARS)	0.035	0.051	0.060	0.087	0.033	0.049	0.054	0.079	0.97	0.97	0.96	0.96	0.97	0.97	0.96	0.96
CLIME (CV-Likelihood)	0.031	0.046	0.058	0.086	0.028	0.043	0.052	0.078	0.77	0.77	0.54	0.54	0.77	0.77	0.55	0.55
CLIME (CV-Trace)	0.032	0.048	0.059	0.087	0.031	0.045	0.053	0.079	0.83	0.83	0.75	0.75	0.83	0.83	0.74	0.74
TIGER (StARS)	0.026	0.040	0.059	0.087	0.023	0.037	0.053	0.079	0.94	0.94	0.96	0.96	0.94	0.94	0.96	0.96
TIGER (CV-Likelihood)	0.032	0.048	0.060	0.088	0.031	0.046	0.054	0.080	0.97	0.97	0.96	0.96	0.97	0.97	0.96	0.96
	TPR								FPR							
	In;Ct		In;Non-ct		Out;Ct		Out;Non-ct		In;Ct		In;Non-ct		Out;Ct		Out;Non-ct	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
FIC - ℓ_2	0.20	0.19	0.17	0.15	0.19	0.19	0.14	0.14	0.08	0.06	0.10	0.07	0.06	0.05	0.07	0.06
FIC - LQA ℓ_1	0.29	0.28	0.20	0.18	0.29	0.29	0.19	0.19	0.11	0.09	0.13	0.10	0.11	0.10	0.12	0.12
GL (CV-Likelihood)	0.89	0.89	0.53	0.53	0.90	0.90	0.50	0.50	0.27	0.27	0.13	0.13	0.25	0.25	0.11	0.11
GL (CV-Trace)	0.75	0.75	0.03	0.03	0.70	0.70	0.02	0.02	0.18	0.18	0.00	0.00	0.15	0.15	0.00	0.00
GL (StARS)	0.67	0.67	0.36	0.36	0.71	0.71	0.40	0.40	0.07	0.07	0.04	0.04	0.08	0.08	0.04	0.04
GL (eBIC)	0.25	0.25	0.00	0.00	0.22	0.22	0.00	0.00	0.03	0.03	0.00	0.00	0.03	0.03	0.00	0.00
CLIME (StARS)	0.23	0.23	0.20	0.20	0.25	0.25	0.23	0.23	0.02	0.02	0.03	0.03	0.02	0.02	0.03	0.03
CLIME (CV-Likelihood)	0.80	0.80	0.76	0.76	0.82	0.82	0.76	0.76	0.19	0.19	0.44	0.44	0.20	0.20	0.43	0.43
CLIME (CV-Trace)	0.68	0.68	0.56	0.56	0.71	0.71	0.60	0.60	0.14	0.14	0.23	0.23	0.14	0.14	0.25	0.25
TIGER (StARS)	0.66	0.66	0.31	0.31	0.70	0.70	0.34	0.34	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03
TIGER (CV-Likelihood)	0.41	0.41	0.27	0.27	0.37	0.37	0.28	0.28	0.01	0.01	0.03	0.03	0.01	0.01	0.03	0.03

Table 2: Simulated data. Empirical mean squared error (MSE), sparsity index (SI), true positive rate (TPR) and false positive rate (FPR) of estimated graphs for 2 focuses, pooled across 300 simulation runs and 60 different simulation settings (3 graph types, 5 scenarios per graph). 'In/Out' refer to the settings where the point was an in-sample or out-of sample point while 'Ct/Non-Ct' refer to settings where the variance at each node was constant or non-constant (randomly sampled at each node).

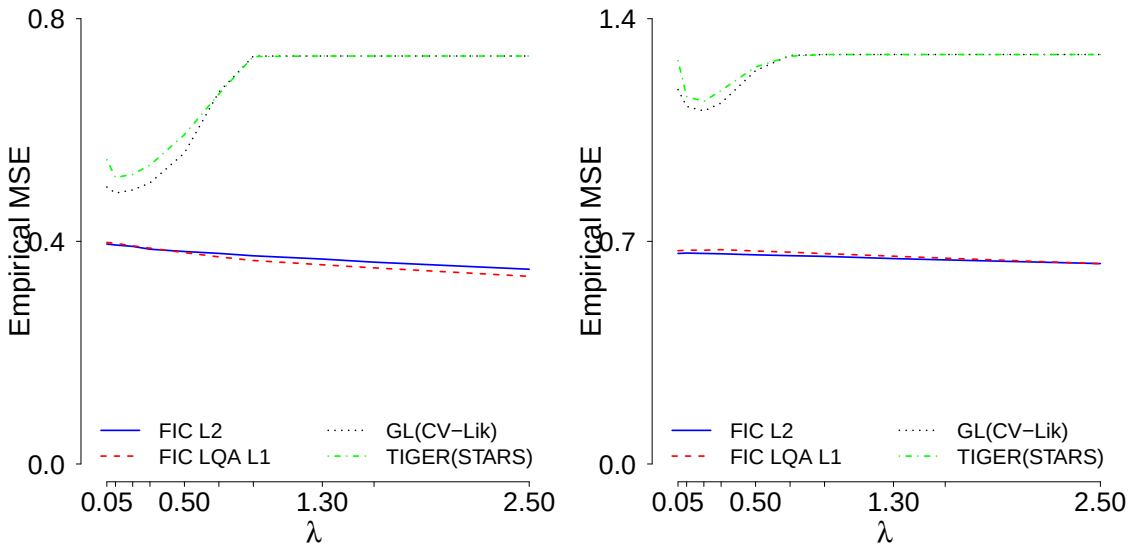


Figure 2: Simulated data. Empirical MSE plotted against a λ sequence for FIC - ℓ_2, FIC - LQA ℓ_1, GL (CV-Likelihood) and TIGER (StARS) when the variance at each node is constant (left panel) or non-constant (right panel) for μ_2 when the evaluation point is out-of-sample.

	In;Ct		In;Non-ct		Out;Ct		Out;Non-ct	
	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2	μ_1	μ_2
FIC - ℓ_2	0.01	0.07	0.01	0.13	0.01	0.07	0.01	0.13
FIC - LQA SCAD	0.01	0.07	0.01	0.13	0.01	0.07	0.01	0.13
TSCGM (eBIC)	0.13	0.13	0.24	0.23	0.15	0.14	0.26	0.25
TSCGM (BIC)	0.13	0.13	0.23	0.24	0.16	0.15	0.27	0.27
TSCGM (GIC)	0.13	0.13	0.24	0.24	0.15	0.14	0.27	0.27

Table 3: Simulated data. Empirical mean squared error (MSE) of estimated graphs for 2 focuses, pooled across 300 simulation runs and 24 different simulation settings (3 graph types, with $n = 30$ or 75 for $p = 35$). 'In/Out' refer to the settings where the point was an in-sample or out-of sample point while 'Ct/Non-Ct' refer to settings where the variance at each node was constant or non-constant (randomly sampled at each node).

coefficients using the undirected graphical structure which accounts for instantaneous relations between nodes corrected for the lagged influence. We explored both ways of getting the empirical MSE values. As expected, using the directed structure provided worse MSE values than using the undirected graph since predicting the nodes by using their past might be too simple as this does not account for the influence of the other nodes. Hence we present the empirical MSE values when using the concentration matrix estimates.

Table 3 presents the empirical MSE values obtained for this comparison. Similar conclusions to independent data hold for the case where autoregressive effects were modeled. The FIC produces smaller empirical MSE values for the focuses, but tends to produce for many settings TPR and FPR rates that are lower than those obtained by using TSCGM. In terms of sparsity, for most settings the FIC produced either comparable or sparser models.

4 Analysis of the fMRI data

For eight participants (2 male, 6 female, mean age= 24.4, range 21–25), resting state functional magnetic resonance imaging data were acquired in a single scanning session on a 3T scanner (Philips). For the resting state protocol participants were instructed to stay alert and focus on a white fixation cross; presented on a black-projection screen that was viewed via a mirror system attached to the magnetic resonance imaging (MRI) head coil. In total, 246 T2*-weighted echoplanar images (EPIs) (2202 mm FOV; 962 in plane resolution; 3.3 mm slice thickness; 0 mm slice spacing; TR 2000 ms; TE 28 ms; FA 90o, ascending orientation) were scanned. For registration purposes, a three-dimensional T1 scan was acquired before functional runs of an independent fMRI study (T1; TFE 218x226 mm FOV; 2562 in plane resolution; 182 slices, 1.2 mm slice thickness, TR 9.56 ms, TE 4.6 ms, FA 8, coronal orientation). FMRI data processing was carried out using FEAT (FMRI Expert Analysis Tool) Version 5.98, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). After discarding the first 6 volumes to allow for stabilization of the magnetic field, the images were concatenated across time into a single 4-dimensional image. The following pre-statistics processing was applied; motion correction using MCFLIRT (Jenkinson et al., 2002); slice-timing correction using Fourier-space time-series phase-shifting; non-brain removal using BET (Smith, 2002); grand-mean intensity normalisation of the entire 4D dataset by a single multiplicative factor; highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma=50.0s). The Lausance 2008 parcellation within the Connectome viewer toolkit (http://www.cmtk.org) was used to create the embedded hierarchical cortical parcellations within Freesurfer (Gerhard et al., 2011; Hagmann et al., 2008; Honey et al., 2009; Cammoun et al., 2012). For each subject, the preprocessed T1-weighted image was first segmented into 68 atlas based cortical parcels (Desikan et al., 2006). See Table 6 for the names of the regions used in the study. In a second step, each region was split into a set of smaller regions on the average space (Freesurfer). The subdivision for each region was then registered to each individual brain in the same way as the original parcellation with 68 regions, such that the regions were similar for each subject. This resulted in a set of 448 regions for each subject. All segmentations were transformed and registered onto the fMRI resting-state images using FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002). Consequently, the averaged times series across voxels was extracted for each of the 448 ROI's, at each time point ($n = 240$). Prior to the computations of networks, all time series were detrended and the mean cerebral fluid and white matter signals were regressed from all time points.

4.1 Focuses of interest

One of the advantages of the FIC is that we can concentrate the graph search on specific parameters (or regions) that will be accurately estimated, that is, have the lowest MSE. We concentrate on three specific focuses concerning the 'default mode network' (DMN), fronto-occipital (FO) regions and the prefrontal cortex (PFC).

μ_1 Regions forming the DMN can be emphasized by having increased levels compared to other regions in the focus. For this focus we set the values of the DMN regions to high or low values and average values for the remaining regions;

μ_2 We emphasize with this focus all ROIs from the fronto-occipital regions by setting the values of these regions to high or low values and average values for the remaining regions;

μ_3 The primary goal of research into executive functions is to investigate the prefrontal cortex. A focus is created that emphasizes the prefrontal regions. This focus obtains high values for brain regions in the PFC and average values for other regions.

The average signal values over the ROIs for focuses μ_1, μ_2 and μ_3 can be seen in Figure 3 in the first column. Each focus represents a certain signal pattern for the ROIs. It is of interest to investigate what the implications are for the structure of the estimated networks. The 'spikes' represent the strength of the average signal in the emphasized regions. For some regions an average strength of the signal is recorded, given by the zero values in Figure 3. For other regions a higher or lower than average signal is recorded. The choice of low/high values for ROIs is based on the observed values for an external subject. The focuses correspond to patterns of activation of a real subject for which the signal in some regions takes an average value, while in some other regions, namely, for μ_1 the DMN, for μ_2 the FO and for μ_3 the PFC, the signal is above or below the average as observed for the external subject.

For the fMRI measurements, we are interested in estimating networks that provide small MSE values for estimating a function of the parameters of interest (focus) $\mu(\theta, \gamma; \tilde{x}) = E(Y|\tilde{x})$.

We observe in Figure 3 that choosing different penalties for a given fixed focus made a smaller difference on the structure of the estimated graphs than choosing a different focus. As expected, the most important factor in determining how the networks look like, is the focus, rather than the penalty. The size of the label for each region is proportional to the corresponding degree of the node and it is comparable across methods, larger labels denote nodes with more connections. For illustration purposes, results for the 3rd subject only are presented throughout this subsection. Results for other subjects are available from the authors.

Comparing the graphs for the same technique, but with different focuses, we see that since the focuses are quite far from each other, the graphs are also identifying different regions as playing central roles. This is to be expected since the FIC procedure optimizes graphs with respect to the focus. In general, the FIC identified the focused regions and produces graphs where these regions are highly connected nodes in the graph. In the FO and PFC focus some regions are common and the FIC is stable since for these regions the graphs share more similarities.

Since the FIC graphs are constructed by optimizing the MSE of the focus estimator we have compared the performance of the graphs presented in Figure 3 with the graphs obtained by using a graphical lasso where we use 3-fold cross-validation with the 'Likelihood' and 'Trace' loss and the StARS criterion for selecting the optimal regularization level on a grid of λ values. The choice for 3-fold cross-validation can be motivated by the assumption of stationary noise, which is reasonable (Wink and Roerdink, 2006). The size of the problem, unfortunately, proved to be too large for comparing to the Bayesian graph and TSGCM techniques. As a function of λ the estimated graphs ranged from being very sparse (they contained a small number of edges) to being very dense (they contained a high number of edges). The same grid was used for the FIC graphs to select the appropriate level of regularization for each method. We present in Table 4 the obtained empirical MSE value for the focused ROIs and the number of edges estimated by each technique as a ratio versus the performance of FIC LQA ℓ_1. For better compatibility with the cross-validation used for GL we chose the regularization level based on the GCV criterion. All three techniques estimate more edges in the graph than the FIC and the stability selection provided an empirical MSE slightly larger than that of the FIC LQA ℓ_1 but at the expense of a much denser graph.

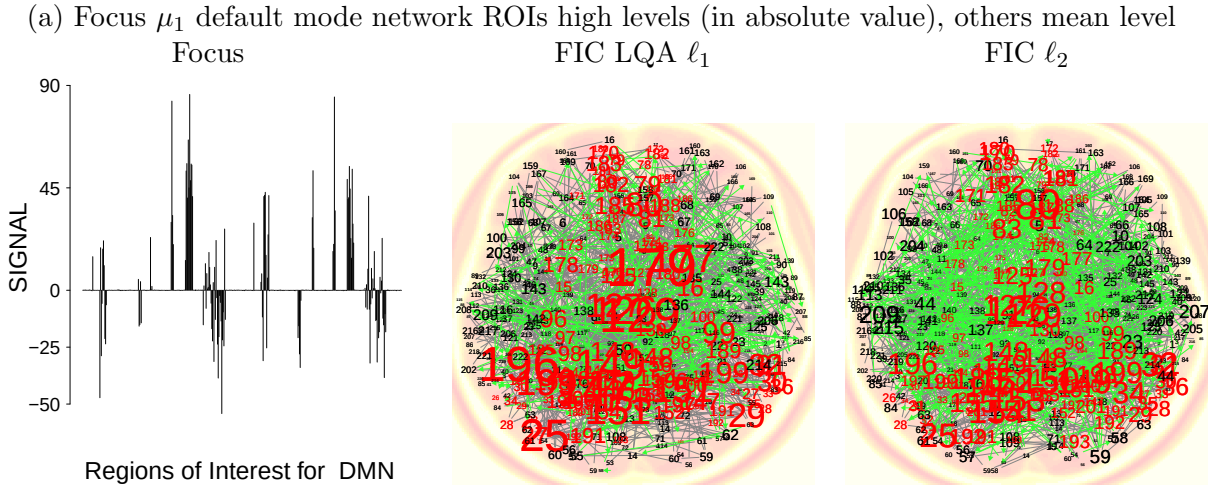
Based on the estimated graphs for the DMN and FO regions we are interested in knowing:

(i) whether there is a propensity of the DMN regions to connect more intensely to other DMN regions or to 'outside' regions, following (Bassett et al., 2008);

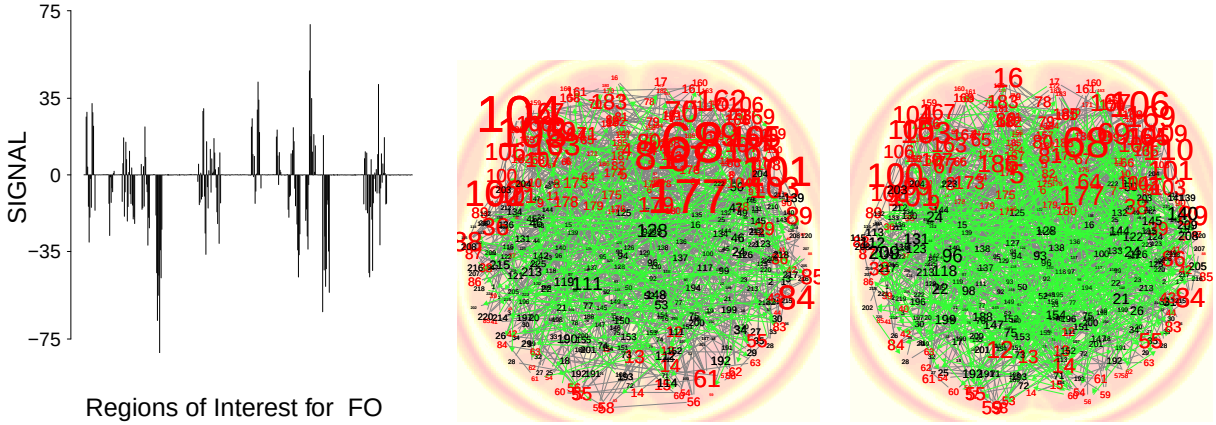
(ii) if the hypothesis of small-worldness and the property of a truncated power-law degree distribution hold for the estimated network based on the fronto-occipital (FO) focus.

For the studied DMN focus presented in Figure 3 (upper row and leftmost panel) the estimated graphs are relatively dense and regardless of the penalty used, roughly one in four connections is a connection between nodes from the DMN regions (see Figure 4 leftmost panel). Most of the DMN

(a) Focus μ_1 default mode network ROIs high levels (in absolute value), others mean level



(b) Focus μ_2 fronto-occipital ROIs high levels (in absolute value), others mean level



(c) Focus μ_3 prefrontal cortex ROIs high levels (in absolute value), others mean level

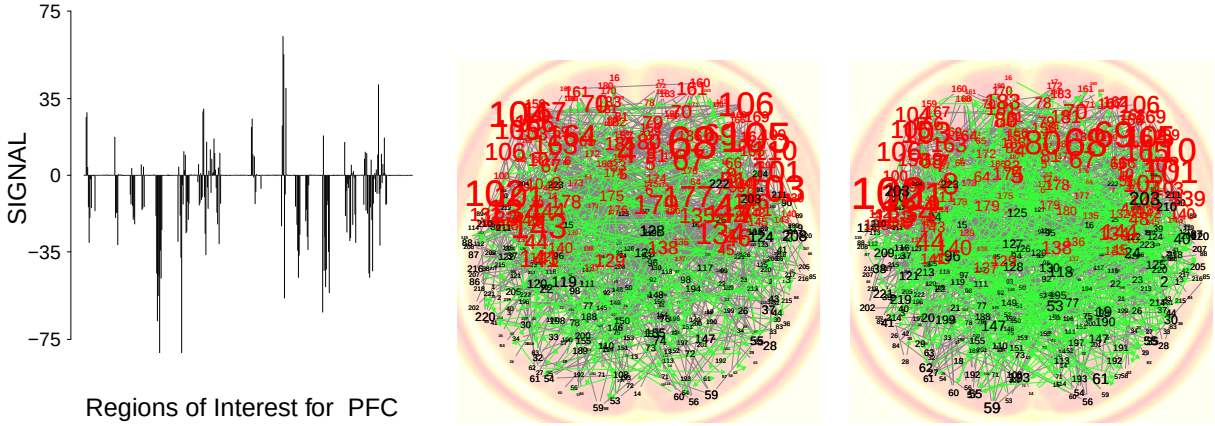


Figure 3: fMRI data. FIC estimated graphs based on three focuses μ_1 – μ_3 for the 3rd subject using penalty functions LQA ℓ_1 and ℓ_2. Undirected edges denote contemporaneous relations while directed edges (bold) denote dynamic relations. Larger labels correspond to ROIs highly connected to other ROIs. Red denotes a ROI from the DMN regions (μ_1), FO regions (μ_2) and PFC regions (μ_3). The signal in the focused regions is higher/lower than the average as specified by the focus given in the first column. Undirected edges are depicted in gray and directed edges are depicted in green.

Ratio vs FIC	GL-CV(Lik)	GL-CV(Trace)	GL-StARS
MSE (μ_1)	1.44	1.44	1.06
MSE (μ_2)	1.36	1.36	1.01
MSE (μ_3)	1.52	1.52	1.11
No of edges (μ_1)	4.40	4.40	9.70
No of edges (μ_2)	4.66	4.66	10.3
No of edges (μ_3)	4.67	4.67	10.3

Table 4: fMRI data. Ratios of the empirical MSE and number of edges in the estimated graphs for GL-CV(Lik), GL-CV(Trace) and GL-StARS relative to the graphs estimated using FIC - LQA ℓ_1 for μ_1, μ_2 and μ_3.

connections are made between ROIs that form the medial temporal lobes. In Figure 4 only the results for FIC - LQA ℓ_1 are shown as the results for FIC ℓ_2 are similar.

Regarding the fronto-occipital regions, the truncated power-law distribution hypothesis posits that the probability of a node to have degree equal to r is proportional to r^{ζ}, where the exponent ζ often ranges from 2 to 3 for biological networks (Bullmore and Sporns, 2009). For the selected graph based on FO regions, the degree distribution seems to be exponentially decaying (see Figure 4, rightmost panel) with rates 2.20 (for FIC - LQA ℓ_1) and 1.99 (for FIC ℓ_2) which makes such a hypothesis plausible.

When compared to an Erdös-Rényi random graph of similar characteristics (the same number of nodes and edges as the observed graph, but the edges are placed with uniform probability), the estimated network appears to have higher clustering, and thus higher local connectivity for roughly the same shortness of paths as in the random case. Thus the hypothesis of small-worldness is justifiable too. This means that most regions can be reached within a few intermediate passes, as almost paradoxically most nodes have a low amount of immediate connections. For this purpose the estimated network was compared to a sample of 10000 random graphs and a histogram of all estimated small-world coefficients is presented in Figure 4. Around 86% (for FIC - LQA ℓ_1) and 99% (for FIC ℓ_2) of the estimated values are larger than the cut-off value of 1 used in (Humphries et al., 2006) but only around 2% (regardless of the penalty) of the values are larger than a conservative value of 3 used in (Humphries and Gurney, 2008).

As a form of validation we have repeated the analysis for all 8 subjects in study (on a smaller dataset containing 68 ROIs, see Table 6 for the names of the regions used) and there is a high degree of reproducibility, in the sense that the regions forming the DMN and the FO get identified as important regions for most of the subjects.

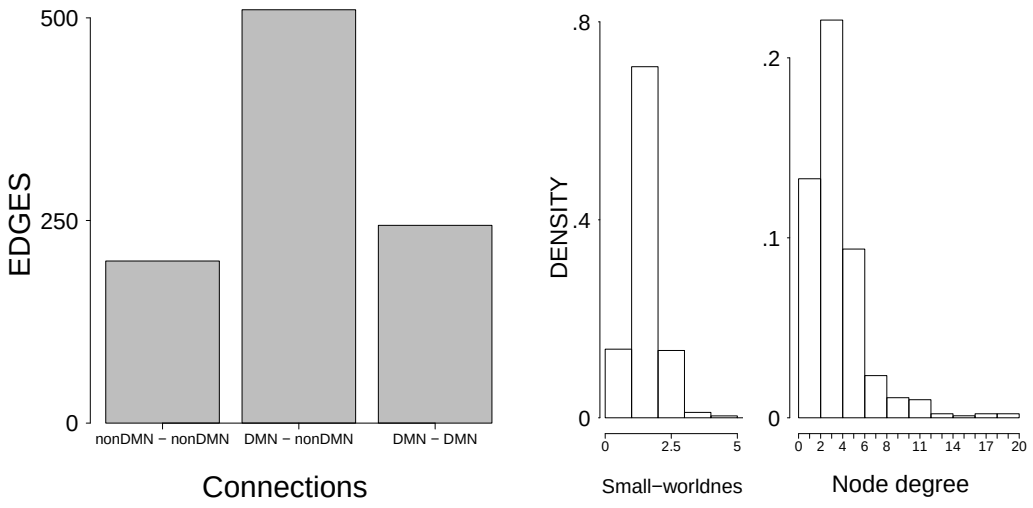


Figure 4: fMRI data. Histograms showing the number of edges connecting the DMN regions (left panel) and the distribution of small-worldness values and node degree for FO regions (right panel). The results for FIC - LQA ℓ_1 are shown.

5 Extensions of the basic model

It is well known that individual differences exist in people's brain structure and function, which may have originated from genetic or phenotypic differences. It seems therefore intuitive that subjects with similar genetic or phenotypic profiles are also similar with respect to brain function or connectivity (Thompson et al., 2001). We can then use the fact that certain connectivity patterns, like hubs, are expected for certain groups of subjects, but not for others. Alternatively, we can incorporate those individual differences explicitly into our model by using a mixed effects model to allow for individual differences between subjects that are drawn from a common population.

5.1 A mixed effects combined-data model

Background information such as genetic markup could lead to assumptions about similar graphs for a class of subjects. It may then be worthwhile to allow for some differences between subjects while still assuming that these subjects are from the same population. Statistically it is then beneficial to use a linear mixed effects model that pools information from all subjects.

We let for this

$$\boldsymbol{Y}_j = \alpha + \boldsymbol{X}_j\beta + \boldsymbol{Z}_j b_j + \epsilon_j,$$

where the index j denotes the subject level with $j = 1, \ldots, 8$ and $\boldsymbol{Y}_j$ represents a vector of n_j measurements; in our case $n_j = 240$ for all 8 subjects in the analysis. The matrices $\boldsymbol{X}_j$ (of dimension $n_j \times p$) and $\boldsymbol{Z}_j$ (of dimension $n_j \times q$) represent the design matrices corresponding to the fixed and random effects. The parameters α and β represent the fixed effects parameters, while b_j represents the vector (of length q) of subject specific effects with $b_j \sim N(0, D)$ where D is the $q \times q$ variance matrix of the random effects. The random errors $\epsilon_j \sim N(0, R_j)$ with R_j a $n_j \times n_j$ variance matrix.

In this application, we model only a random intercept and therefore we let $\boldsymbol{Z}_j$ denote a vector of ones (of length n_j), though the structure of $\boldsymbol{Z}_j$ can be more complex. We treat all dynamic and contemporaneous effects as fixed, but one could, if desired, treat any of them as random by including the observed measurements corresponding to these nodes in the $\boldsymbol{Z}_j$ matrix. Here it is assumed that $D = \sigma_b^2 I_q$ and $R_j = \sigma^2 I_{n_j}$ for all j which implies that for all subjects σ^2 and σ_b^2 are constant.

In light of the notation in Section 2.2 we have $\theta = (\sigma_b^2, \sigma^2, \alpha)$ and $\gamma = \beta$ and the complete loglikelihood with a penalty function on the β coefficients is used in the optimization problem:

$$\begin{aligned} Q(\theta,\gamma) &= -\frac{1}{2}\sum_{j=1}^{8}\Big[\log\det\Sigma_j + \Big\{\begin{pmatrix}\boldsymbol{Y}_j^{\mathsf{T}}\\ b_j^{\mathsf{T}}\end{pmatrix} - \begin{pmatrix}\alpha+\boldsymbol{X}_j\beta\\ 0\end{pmatrix}\Big\}^{\mathsf{T}}\Sigma_j^{-1}\Big\{\begin{pmatrix}\boldsymbol{Y}_j^{\mathsf{T}}\\ b_j^{\mathsf{T}}\end{pmatrix} - \begin{pmatrix}\alpha+\boldsymbol{X}_j\beta\\ 0\end{pmatrix}\Big\}\Big] \\ &\quad -\lambda\sum_{l=1}^{p}\psi(|\beta_l - \beta_{l0}|) \text{ where } \Sigma_j = \begin{pmatrix}\sigma_b^2\boldsymbol{Z}_j\boldsymbol{Z}_j^{\mathsf{T}} + \sigma^2 I_{n_j} & \sigma_b^2\boldsymbol{Z}_j\\ \sigma_b^2\boldsymbol{Z}_j^{\mathsf{T}} & \sigma_b^2 I_q\end{pmatrix}. \end{aligned}$$

To estimate the unknown parameters we use the EM algorithm (Laird et al., 1987). For our example with 8 subjects and $n_j = 240$ time points, similar to the algorithm in McLachlan and Krishnan (2008), we set at iteration (k)

$$b_j^{(k)} = \Big(\boldsymbol{Z}_j^{\mathsf{T}}\boldsymbol{Z}_j + I_q\sigma^{2^{(k)}}\sigma_b^{-2^{(k)}}\Big)^{-1}\boldsymbol{Z}_j^{\mathsf{T}}(\boldsymbol{Y}_j - \alpha^{(k)} - \boldsymbol{X}_j\beta^{(k)}),$$

where $P = \text{diag}(\psi'(|\beta^{(k)}|)/|\beta^{(k)}|)$ and at iteration $(k+1)$ we set

$$\begin{aligned} \begin{pmatrix}\alpha\\ \beta\end{pmatrix}^{(k+1)} &= \Big(\sum_{j=1}^{8}([1_{n_j}, \boldsymbol{X}_j]^{\mathsf{T}}[1_{n_j}, \boldsymbol{X}_j] + \lambda P\Big)^{-1}\sum_{j=1}^{8}([1_{n_j}, \boldsymbol{X}_j]^{\mathsf{T}}(\boldsymbol{Y}_j - \boldsymbol{Z}_j b_j^{(k)}) \\ \sigma^{2^{(k+1)}} &= \frac{1}{240\cdot 8}\sum_{j=1}^{8}[\text{trace}\{\boldsymbol{Z}_j^{\mathsf{T}}\boldsymbol{Z}_j(\sigma^{-2^{(k)}}\boldsymbol{Z}_j^{\mathsf{T}}\boldsymbol{Z}_j + \sigma_b^{-2^{(k)}}I_q)\} \\ &\quad + (\boldsymbol{Y}_j - \alpha^{(k)} - \boldsymbol{X}_j\beta^{(k)} - \boldsymbol{Z}_j b_j^{(k)})^{\mathsf{T}}(\boldsymbol{Y}_j - \alpha^{(k)} - \boldsymbol{X}_j\beta^{(k)} - \boldsymbol{Z}_j b_j^{(k)})] \\ \sigma_b^{2^{(k+1)}} &= \frac{1}{240\cdot 8}\sum_{j=1}^{8}[\text{trace}\{(\sigma^{2^{(k)}} + \sigma_b^{2^{(k)}})^{-1}\sigma^{2^{(k)}}\sigma_b^{2(k)}I_q\} + b_j^{(k)\mathsf{T}}b_j^{(k)}]. \end{aligned}$$

The simplest model that we consider for a node, does not include any other nodes as neighbors (i.e $\gamma_0 = 0$), and the most complex model includes all other nodes as potential neighbors.

Let $\tilde{Y}$ represent the stacked vector of measurements from all 8 subjects i.e. $\tilde{Y} = (\boldsymbol{Y}_1^{\mathsf{T}}, \ldots, \boldsymbol{Y}_8^{\mathsf{T}})^{\mathsf{T}} = (Y_{1,1}, \ldots, Y_{1,240}, \ldots, Y_{8,1}, \ldots, Y_{8,240})^{\mathsf{T}}$ and let $\tilde{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_8)$ be the design matrix corresponding to the stacked fixed effects design matrices. Let $\tilde{Z} = \text{Diag}\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_8\}$ be a diagonal design matrix where the individual design matrices for each subject are placed on the main diagonal and let $\tilde{V}$ be a diagonal matrix constructed as $\tilde{V} = \text{Diag}\{\sigma_b^2\boldsymbol{Z}_1\boldsymbol{Z}_1^{\mathsf{T}} + \sigma^2 I_{n_1}, \ldots, \sigma_b^2\boldsymbol{Z}_8\boldsymbol{Z}_8^{\mathsf{T}} + \sigma^2 I_{n_8}\}$.

The Fisher information matrix J takes for this particular case the following form

$$J = \text{Diag}\{J_{00},\ \tilde{X}^{\mathsf{T}}\tilde{V}^{-1}\tilde{X} + \lambda P\},$$

where the submatrix corresponding to the elements of θ,

$$J_{00} = \text{Diag}\left\{\begin{pmatrix}\frac{1}{2}\text{trace}\{\tilde{V}^{-1}\frac{\partial\tilde{V}}{\partial\sigma_b^2}\tilde{V}^{-1}\frac{\partial\tilde{V}}{\partial\sigma_b^2}\} & \frac{1}{2}\text{trace}\{\tilde{V}^{-1}\frac{\partial\tilde{V}}{\partial\sigma_b^2}\tilde{V}^{-1}\frac{\partial\tilde{V}}{\partial\sigma^2}\}\\ \frac{1}{2}\text{trace}\{\tilde{V}^{-1}\frac{\partial\tilde{V}}{\partial\sigma^2}\tilde{V}^{-1}\frac{\partial\tilde{V}}{\partial\sigma_b^2}\} & \frac{1}{2}\text{trace}\{\tilde{V}^{-1}\frac{\partial\tilde{V}}{\partial\sigma^2}\tilde{V}^{-1}\frac{\partial\tilde{V}}{\partial\sigma^2}\}\end{pmatrix},\ 1_{8\cdot 240}^{\mathsf{T}}\tilde{V}^{-1}1_{8\cdot 240}\right\},$$

where $\frac{\partial\tilde{V}}{\partial\sigma_b^2} = \text{Diag}\{\boldsymbol{Z}_1\boldsymbol{Z}_1^{\mathsf{T}}, \ldots, \boldsymbol{Z}_8\boldsymbol{Z}_8^{\mathsf{T}}\}$, $\frac{\partial\tilde{V}}{\partial\sigma^2} = \text{Diag}\{R_1, \ldots, R_8\}$ and $1_{8\cdot 240}$ is a vector of ones of the same length as the number of rows of $\tilde{V}^{-1}$.

Model selection for the pooled data follows the same steps as in Section 4.1. Under the above framework, we estimate the vectors $\hat{\theta}$ and $\hat{\gamma}$ using the largest model. We then construct the empirical version of the Fisher information matrix, partition it according to the lengths of the vectors θ and γ, construct all the necessary quantities using these partitions, the partial derivatives

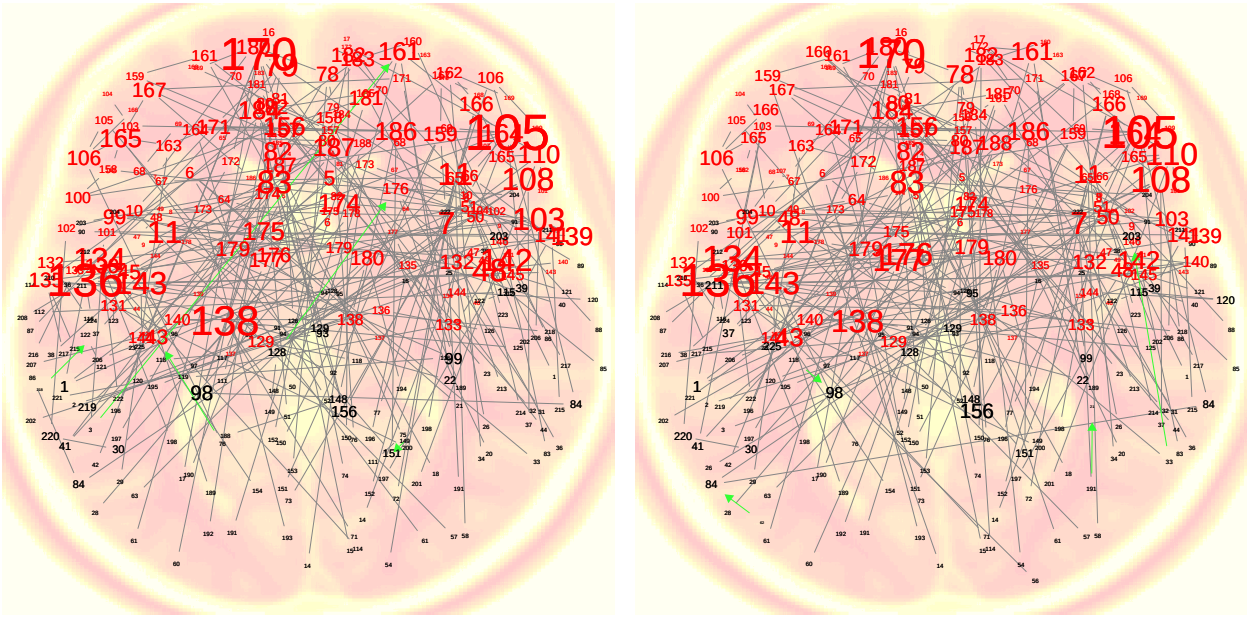


Figure 5: fMRI data. FIC estimated graphs with pooled data and subject specific effects based on μ_3 and using the penalty functions LQA ℓ_1 (left panel) and ℓ_2 (right panel). We refer to Figure 3 for details about what the colors and edges represent.

of the focuses and the projection matrices for each model and in the end calculate the MSE value as in (A.1). We use the same algorithms as in Section 2.6. Note that the purpose of this extension is to allow for the estimation of a general graph when the information from all subjects is pooled together. The differences between subjects are captured when estimating the fixed (common to all subjects) and random (subject specific) parameters. This is different from estimating subject specific networks as in Section 4. The goal is here to pool all data in order to estimate one graph.

Figure 5 presents the obtained results when FIC is used on the pooled data when different subject-specific effects are allowed in the models at each node and when the ℓ_2 and LQA ℓ_1 penalties are used for constructing the P matrix. We consider the same PFC focus as in Section 4.1. Comparing the obtained results with those of Section 4.1 we conclude that the ROIs with large (in absolute value) signal values get selected as important regions and that pooling information from all subjects results in having estimated sparser graphs. Again, the focus is more important in determining the structure of the graph than the penalty used.

5.2 An average model

By specifying the time as part of the focus, FIC is able to estimate different graphs for each time-point. An averaged graph is constructed where averaging takes place over all 8 subjects and all time points. There is some evidence that the functional graph changes over time, possibly reflecting different states Cribben et al. (2012). Our average graph should therefore be interpreted as reflecting common properties over subjects and stable edges in the graph. We compared the FIC with a GL approach where we choose the graph according to the stability selection procedure (GL StARS) and cross-validation (GL CV-Likelihood). For FIC we have used an ℓ_2, an LQA ℓ_1 and the LQA SCAD penalty. Table 5 presents the obtained results averaged across all 8 subjects.

With respect to the transitivity (Tran.) of the estimated networks, the FIC ℓ_2 graphs were closer to the GL graphs than the other penalties. This means that if two ROIs communicate with a third one, the FIC ℓ_2 and GL would suggest more often that they should influence each other as well, thus suggesting more clustered brain regions than the other graphs.

The average degree (Avg. Dgr.) and the maximal degree (Max. Dgr.) of a node tended to be higher for FIC graphs than for the GL graphs. This suggests that the FIC graphs have on average more edges connecting a node, thus a ROI will have more connections linking it to other regions in the FIC graph. Moreover, the average stress centrality score (Avg. St. Cnt.) suggests that for the FIC - LQA ℓ_1 and LQA SCAD there are fewer shortest paths between other ROIs passing through a specific node than for GL. This means that going from one region to another can be accomplished in fewer ways once the information passes a hub. The average path length (Avg. Pth.) between two nodes was also generally larger for the FIC graphs than for the GL graphs, showing that on average the shortest path between two nodes is larger in the FIC graph which implies that for these graphs the information would travel longer.

The average betweennees values (Avg. Btw.) suggest that in the FIC estimated models, information can flow from one ROI to another one on several distinct routes that pass through a 'gateway', while for the GL graphs, this number is smaller and thus on average moving between ROIs can be accomplished in fewer ways through a specific node.

A small-world (SWI) behavior is not strongly supported when using the FIC - LQA ℓ_1 and LQA SCAD graphs, as on average the SWI values are close to the cut-off value of 1, but the estimated graphs using FIC ℓ_2 support such a hypothesis which is in line with claims as in (Sporns and Honey, 2006) or Achard et al. (2006).

Another common feature of the network summary statistics are the positive assortativity coefficients (Asso.). This implies a preference for nodes to attach to others that are more similar in terms of degree. Thus, high degree nodes are linked with other high degree nodes, and low degree nodes are more often linked with other low degree nodes, but the relation is not very strong. The number of estimated clusters (Max. Kcor.) suggests that the number of groups of nodes which are disconnected from other groups of nodes is similar for FIC graphs and GL graphs.

We further construct an 'average' model by retaining only the edges that appear with a high frequency in such a way that the sparsity of the FIC network was set close to that of the GL StARS. Figure 6 presents the estimated average networks for the 3rd subject, using FIC - LQA ℓ_1 and the graph estimated with GL StARS. Visually, both networks seem to roughly indicate similar structures of interaction between ROIs. Most often, the graphs proposed the 8th (inferiorparietal), 10th (insula), 20th (parsorbitalis), 23rd (postcentral), 24th (posteriorcingulate), 25th (precentral), 28th (rostralmiddlefrontal), 29th (superiorfrontal) and 32nd (supramarginal) ROI as important regions in brain functionality. Both estimated graphs agree in finding these regions as important ones, but vary somewhat in the number of links connecting these regions to the other ones.

Procedure	Tran.	Avg. Pth	Avg. Dgr.	Max. Dgr.	Avg. St.Cnt.	Avg. Btw.	Asso.	SWI	Max. Kcor.
FIC - ℓ_2	0.37	3.42	11.57	19.50	596.00	167.73	0.17	1.42	8.12
FIC - LQA ℓ_1	0.21	2.51	12.64	21.25	362.88	115.79	0.04	0.94	8.62
FIC - LQA SCAD	0.21	2.51	12.63	21.00	362.38	115.75	0.04	0.93	8.62
GL CV-Likelihood	0.38	2.28	9.86	19.88	483.00	42.55	0.18	2.38	13.50
GL StARS	0.38	2.25	9.99	19.75	367.38	41.53	0.16	2.23	9.25

Table 5: fMRI data. Network summary measures for estimated mixed graphs including contemporaneous and dynamic effects for eight subjects.

(Moussa et al., 2012) concluded based on the results of a voxel-based analysis that the precentral/postcentral regions are consistent regions in the sensory/motor module across subjects, while (Koyama et al., 2011) found that high functional connectivity between the precentral ROI and other motor areas was positively correlated with reading abilities. (Fan et al., 2012) discovered a decreased activity in the postcentral gyrus (among a few others) for adults with depression episodes of bipolar disorder. There is, thus, already some evidence in the literature which seems to confirm

(a) FIC LQA ℓ_1 (b) GL-StARS

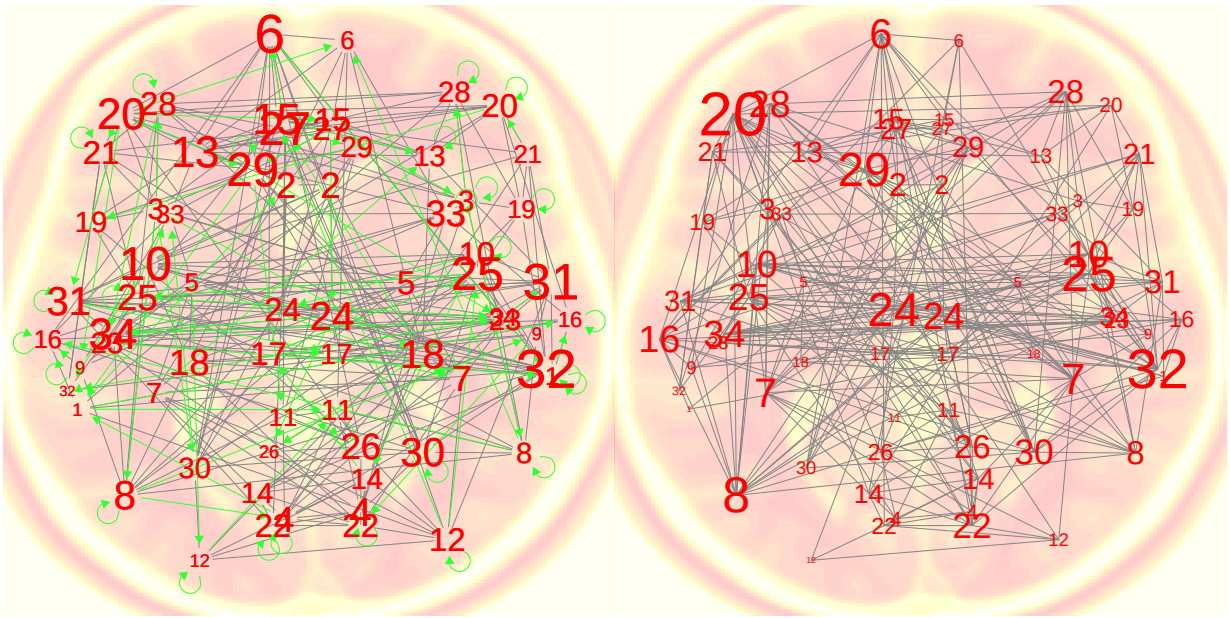


Figure 6: fMRI data. Estimated graphical structures for the 3rd subject using FIC - LQA ℓ_1 (a) and graphical lasso with stability selection (b). Larger labels correspond to high-degree ROIs. Undirected edges denote contemporaneous effects while directed edges denote dynamic effects. The graphical lasso models only contemporaneous effects.

that the regions identified by FIC have a theoretical basis for their presence in the estimated graphs.

6 Discussion

Selecting graphs using the FIC can be a fruitful modeling strategy, as a direct estimation of the MSE of a focus estimator is performed which incorporates the research interest. Focuses based on different configurations of covariates can lead to different selected graphical models and moreover using different μ functions for the same configuration of covariates can lead too to different models being selected. This is not a methodological contradiction as all of the situations above relate to different research questions, which should not necessarily receive the same answer. The FIC offers thus more flexibility in selecting models, and orients the search toward a particular interest. In contrast, all competitors used in this study result in only one selected model regardless of the specific purpose for which model selection is desired. With FIC one can extract more specific information from the analysis. The analysis on the fMRI dataset revealed potential configurations of edges where some regions of the brain are revealed as important regions acting as informational hubs where individual particularities and differences in signal patterns can be easily identified and assessed. The performance on simulated datasets showed that model selection based on FIC can be a powerful and beneficial strategy.

The accumulation of knowledge about connectivity from both animal and human research leads to the possibility of informed analyses of connectivity. This has the great advantage that (anatomical) information about regions of interest or connectivity can be taken into account by making a focus. For large-scale networks such an approach is novel. A focus is an informed way of describing the importance of the prior knowledge. Using the FIC to determine which network has the lowest estimated MSE results in high accuracy for the choice of the parameters in focus. As was seen in resting state analysis, the focus on the prefrontal cortex results in an emphasis on connectivity in

the prefrontal cortex.

The implications for fMRI research are clear. (i) Prior research on connectivity is explicitly taken into account by the use of a focus. Commonly, a meta-analysis is used to see what the current position of the field is on, say, the prefrontal cortex (Ridderinkhof et al., 2004). Due to focusing on a particular region or set of pathways, a new study reflects both prior and current research. With the FIC prior research is not hard-coded into the algorithm, but is used to emphasize in the current estimates what has been found before. (ii) Accuracy (in terms of MSE) of the focus is high, resulting in a network that is tailored to the specific needs of the researcher. This may especially prove relevant for prediction. When the focus can be related to behavioral data, for instance, then prediction will be more accurate using the model that optimizes the FIC than one that optimizes the BIC, say. This is because the FIC balances the squared bias and variance of the focus such that the MSE of the focus estimator is minimal. This in turn implies that prediction with the network obtained with the FIC is optimal with respect to the focus.

A Details of calculations

A.1 Notation in submodels and definition of the FIC

We denote by J the Fisher information matrix of the full model evaluated at the narrow model parameter (θ_0, γ_0) with the inverse matrix denoted by J^{-1} and we define J_S to be the Fisher information matrix of model S, including only those rows and columns indexed by S. All three matrices are partitioned to the dimensions of θ and γ as $J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix}$, $J^{-1} = \begin{pmatrix} J^{00} & J^{01} \\ J^{10} & J^{11} \end{pmatrix}$ and $J_S = \begin{pmatrix} J_{00,S} & J_{01,S} \\ J_{10,S} & J_{11,S} \end{pmatrix}$. We further introduce the following quantities: $\omega = J_{10}J_{00}^{-1}\frac{\partial \mu}{\partial \theta} - \frac{\partial \mu}{\partial \gamma}$, $G_S = J^{11,S,0}\left(J^{11}\right)^{-1}$, $J^{11,S,0} = \pi_S^{\mathsf{T}} J^{11,S}\pi_S$ and $c_n = \lambda\psi''(0)1_q/\sqrt{n} \to c$, where π_S is a projection matrix containing only 0's and 1's corresponding to selecting only those components indicated by S and 1_q is a vector of 1's of length q. For any model S we have (see Claeskens, 2012)

$$\sqrt{n}(\hat{\mu}_S - \mu_{\text{true}}) \xrightarrow{\mathcal{D}} \Lambda_S \sim N\left(\omega^{\mathsf{T}}((I - G_S)\delta - J^{11,S,0}c), \left(\frac{\partial \mu}{\partial \theta}\right)^{\mathsf{T}} J_{00}^{-1}\frac{\partial \mu}{\partial \theta} + \omega^{\mathsf{T}} J^{11,S,0}\omega\right),$$

$$\begin{aligned}
\text{MSE}(\hat{\mu}_S) &= \left(\frac{\partial \mu}{\partial \theta}\right)^{\mathsf{T}} J_{00}^{-1}\frac{\partial \mu}{\partial \theta} + \omega^{\mathsf{T}} J^{11,S,0}\omega + \omega^t((I - G_S)\delta - J^{11,S,0}c)((I - G_S)\delta - J^{11,S,0}c)^{\mathsf{T}}\omega \\
&= \left(\frac{\partial \mu}{\partial \theta}\right)^{\mathsf{T}} J_{00}^{-1}\frac{\partial \mu}{\partial \theta} + \omega^{\mathsf{T}} J^{11,S,0}\omega + \omega^{\mathsf{T}}\{(I - G_S)\delta\delta^{\mathsf{T}}(I - G_S^{\mathsf{T}})\}\omega + \\
&\quad + \omega^{\mathsf{T}}\{J^{11,S,0}cc^{\mathsf{T}}(J^{11,S,0})^t - 2(I - G_S)\delta c^{\mathsf{T}}(J^{11,S,0})^{\mathsf{T}}\}\omega. \qquad \text{(A.1)}
\end{aligned}$$

A.2 Minimum MSE regularization level

Performing multiplications and leaving out terms that do not depend on c, minimizing (A.1) is equivalent to

$$\min_c \left(\omega^{\mathsf{T}}\{J^{11,S,0}cc^{\mathsf{T}}(J^{11,S,0})^t - 2(I - G_S)\delta c^{\mathsf{T}}(J^{11,S,0})^{\mathsf{T}}\}\omega\right).$$

Taking the derivative of the MSE expression with respect to c and setting it equal to 0, we get the following equation with $c_n = \lambda\psi''(0)1_q/\sqrt{n}$ replacing c,

$$\frac{\lambda\psi''(0)}{\sqrt{n}}\omega^{\mathsf{T}} J^{11,S,0}1_q 1_q{}^{\mathsf{T}}(J^{11,S,0})^{\mathsf{T}}\omega - \omega^{\mathsf{T}}(I - G_S)\delta 1_q{}^{\mathsf{T}}(J^{11,S,0})^{\mathsf{T}}\omega = 0,$$

which is solved by

$$\lambda_S = \frac{\omega^\mathsf{T}(I - G_S)\delta 1_q{}^\mathsf{T}(J^{11,S,0})^\mathsf{T}\omega}{\omega^\mathsf{T}J^{11,S,0}1_q 1_q{}^\mathsf{T}(J^{11,S,0})^\mathsf{T}\omega}\frac{\sqrt{n}}{\psi''(0)}. \tag{A.2}$$

B Correspondence between the studied ROIs and the numbering

ROI	Name	ROI	Name	ROI	Name
1	Bankssts	12	Lateraloccipital	23	Postcentral
2	Caudalanteriorcingulate	13	Lateralorbitofrontal	24	Posteriorcingulate
3	Caudalmiddlefrontal	14	Lingual	25	Precentral
4	Cuneus	15	Medialorbitofrontal	26	Precuneus
5	Entorhinal	16	Middletemporal	27	Rostralanteriorcingulate
6	Frontalpole	17	Paracentral	28	Rostralmiddlefrontal
7	Fusiform	18	Parahippocampal	29	Superiorfrontal
8	Inferiorparietal	19	Parsopercularis	30	Superiorparietal
9	Inferiortemporal	20	Parsorbitalis	31	Superiortemporal
10	Insula	21	Parstriangularis	32	Supramarginal
11	Isthmuscingulate	22	Pericalcarine	33	Temporalpole
				34	Transversetemporal

Table 6: fMRI data. Correspondence between numbers and names of the regions of interest.

Acknowledgements

We are thankful to the editor and all five reviewers of this manuscript for their questions and suggestions for improvement. We acknowledge the support of the Fund for Scientific Research Flanders, KU Leuven grant GOA/12/14 and of the IAP Research Network P7/06 of the Belgian Science Policy The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government - department EWI.

References

Abegaz, F. and Wit, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, 14(3):586–599.

Achard, S., Salvador, R., Whitcher, B., Suckling, J., and Bullmore, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *The Journal of Neuroscience*, 26(1):63–72.

Allen, E., Damaraju, E., Plis, S., Erhardt, E., Eichele, T., and Calhoun, V. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex*, 24(3):663–676.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516.

Bassett, D. S., Bullmore, E., Verchinski, B. A., Mattay, V. S., Weinberger, D. R., and Meyer-Lindenberg, A. (2008). Hierarchical organization of human cortical networks in health and schizophrenia. *The Journal of Neuroscience*, 28(37):9239–9248.

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences*, 1124(1):1–38.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198.

Bunea, F., She, Y., Ombao, H., Gongvatana, A., Devlin, K., and Cohen, R. (2011). Penalized least squares regression methods and applications to neuroimaging. *Neuroimage*, 55(4):1519–1527.

Cai, T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Cammoun, L., Gigandet, X., Meskaldji, D., Thiran, J. P., Sporns, O., Do, K. Q., Maeder, P., Meuli, R., and Hagmann, P. (2012). Mapping the human connectome at multiple scales with diffusion spectrum mri. *Journal of neuroscience methods*, 203(2):386–397.

Chai, X. J., Whitfield-Gabrieli, S., Shinn, A. K., Gabrieli, J. D., Nieto Castañón, A., McCarthy, J. M., Cohen, B. M., and Ongür, D. (2011). Abnormal medial prefrontal cortex resting-state connectivity in bipolar disorder and schizophrenia. *Neuropsychopharmacology*, 36(10):2009–2017.

Claeskens, G. (2012). Focused estimation and model averaging with penalization methods, an overview. *Statistica Neerlandica*, 66(3):272–287.

Claeskens, G. and Hjort, N. (2003). The focused information criterion. *Journal of the American Statistical Association*, 98:900–916.

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403.

Cribben, I., Haraldsdottir, R., Atlas, L., Wager, T., and Lindquist, M. (2012). Dynamic connectivity regression: Determining state-related changes in brain connectivity. *NeuroImage*, 61(4):907 – 920.

Dahlhaus, R. and Eichler, M. (2003). Causality and graphical models in time series analysis. *Oxford Statistical Science Series*, pages 115–137.

Dempster, A. (1972). Covariance selection. *Biometrics*, 28(1):157–175.

Deshpande, G., Santhanam, P., and Hu, X. (2011). Instantaneous and causal connectivity in resting state brain networks derived from functional MRI data. *Neuroimage*, 54(2):1043–1052.

Desikan, R. S., Sè gonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., and Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31(3):968 – 980.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521–541.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, T., Yao, L., and Wu, X. (2012). Independent component analysis of the resting-state brain functional MRI study in adults with bipolar depression. In *Proceedings of 2012 International Conference on Complex Medical Engineering*, pages 38–42.

Foygel, R. and Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 604–612. (NIPS).

Frank, M. (2011). Computational models of motivated action selection in corticostriatal circuits. *Current Opinion in Neurobiology*, 21:381–386.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friston, K. J., Kahan, J., Biswal, B., and Razi, A. (2014). A DCM for resting state fMRI. *Neuroimage*, 94:396 – 407.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416.

Gao, W. and Tian, Z. (2010). Latent ancestral graph of structure vector autoregressive models. *Journal of Systems Engineering and Electronics*, 21(3):233–238.

Gerhard, S., Daducci, A., Lemkaddem, A., Meuli, R., Thiran, J. P., and Hagmann, P. (2011). The connectome viewer toolkit: an open source framework to manage, analyze, and visualize connectomes. *Frontiers in neuroinformatics*, 5:3.

Hagmann, P. Cammoun, L., Gigandet, X., Meuli, R., Honey, C., Wedeen, J., and Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biol*, 6(7):e159.

Honey, C. J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J. P., Meuli, R., and Hagmann, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences*, 106(6):2035–2040.

Humphries, M., Gurney, K., and Prescott, T. (2006). The brainstem reticular formation is a small-world, not scale-free, network. *Proceedings of the Royal Society B*, 273:503–511.

Humphries, M. D. and Gurney, K. (2008). Network 'small-world-ness': A quantitative method for determining canonical network equivalence. *PLoS ONE*, 3(4):e0002051.

Hunter, R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics*, 33(4):1617–1642.

Isoda, M. and Hikosaka, O. (2007). Switching from automatic to controlled action by monkey medial frontal cortex. *Nature Neuroscience*, 10(2):240–248.

Jahfari, S., Verbruggen, F., Frank, M. J., Waldorp, L. J., Colzato, L., Ridderinkhof, K. R., and Forstmann, B. U. (2012). How preparation changes the need for top-down control of the basal ganglia when inhibiting premature actions. *The Journal of Neuroscience*, 32:10870–10878.

Jahfari, S., Waldorp, L., van den Wildenberg, W., Scholte, H., Ridderinkhof, K., and Forstmann, B. (2011). Effective connectivity reveals important roles for both the hyperdirect (fronto-subthalamic) and the indirect (fronto-striatal-pallidal) fronto-basal ganglia pathways during response inhibition. *The Journal of Neuroscience*, 31:6891–6899.

James, G. A., Kelley, M. E., Craddock, R. C., Holtzheimer, P. E., Dunlop, B., Nemeroff, C., and Hu, X. P. (2009). Exploratory structural equation modeling of resting-state fMRI: applicability of group models to individual subjects. *Neuroimage*, 45(3):778–787.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379. University of California Press.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841.

Jenkinson, M. and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156.

Kolar, M., Song, L., Ahmed, A., and Xing, E. P. (2010). Estimating time-varying networks. *The Annals of Applied Statistics*, 4(1):94–123.

Koyama, M., Di Martino, A., Zuo, X., Kelly, C., Mennes, M., Jutagir, D., Castellanos, F., and Milham, M. (2011). Resting-state functional connectivity indexes reading competence in children and adults. *The Journal of Neuroscience*, 31(23):8617–8624.

Krishnamurthy, V., Ahipaşaoğlu, S., and d'Aspremont, A. (2012). A pathwise algorithm for covariance selection. In Sra, S., Nowozin, S., and Wright, S., editors, *Optimization for Machine Learning*, pages 479–494. MIT Press.

Laird, N. M., Lange, N., and Stram, D. (1987). Maximum likelihood computation with repeated measures. *Journal of the American Statistical Association*, 83:97–105.

Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.

Lei, Y., Tong, L., and Yan, B. (2013). A mixed l2 norm regularized HRF estimation method for rapid event-related fMRI experiments. *Computational and Mathematical Methods in Medicine*, 2013.

Leonardi, N., Richiardi, J., Gschwind, M., Simioni, S., Annoni, J. M., Schluep, M., and Van De Ville, D. (2013). Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest. *NeuroImage*, 83:937 – 950.

Li, L. and Toh, K. C. (2010). An inexact interior point method for l_1-regularized sparse covariance selection. *Mathematical Programming Computation*, 2(3-4):291–315.

Li, X., Zhao, T., and Liu, H. (2013). *camel: Calibrated machine learning*. R package version 0.2.0.

Lindquist, M. (2008). The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–566.

Liu, H. and Wang, L. (2012). Tiger: A tuning-insensitive approach for optimally estimating large undirected graphs. Technical.

Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125–2149.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley, 2. edition.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Mohammadi, A. and Wit, E. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138.

Moussa, M., Steen, M., Laurienti, P., and Hayasaka, S. (2012). Consistency of network modules in resting-state fMRI connectome data. *PLoS ONE*, 7(8):e44428.

O'Neil, E., Hutchison, R., McLean, D., and S, K. (2014). Resting-state fMRI reveals functional connectivity between face-selective perirhinal cortex and the fusiform face area related to face inversion. *Neuroimage*, 92:349–355.

Pircalabelu, E., Claeskens, G., and Waldorp, L. (2015). A focused information criterion for graphical models. *Statistics and Computing.*, page In press.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2):676–682.

Ravikumar, P. D., Raskutti, G., Wainwright, M. J., and Yu, B. (2008). Model selection in Gaussian graphical models: High-dimensional consistency of l_1-regularized MLE. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, pages 1329–1336. (NIPS).

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., and Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, 306(5695):443–447.

Ryali, S., Chen, T., Supekar, K., and Menon, V. (2012). Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage*, 59(4):3852–3861.

Ryali, S., Supekar, K., Abrams, D., and V., M. (2010). Sparse logistic regression for whole-brain classification of fMRI data. *Neuroimage*, 51(2):752–764.

Scheinberg, K. and Rish, I. (2010). Learning sparse Gaussian Markov networks using a greedy coordinate ascent approach. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part III*, pages 196–212.

Schmidt, M., Niculescu-Mizil, A., and Murphy, K. (2007). Learning graphical model structure using ℓ_1-regularization paths. In *Proceedings of the 22nd national conference on Artificial intelligence*, volume 2, pages 1278–1283. AAAI Press.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155.

Sporns, O. and Honey, C. (2006). Small worlds inside big brains. *Proceedings of the National Academy of Sciences*, 103(51):19219–19220.

Thompson, P. M., Cannon, T. D., Narr, K. L., van Erp, T., Poutanen, V. P., Huttunen, M., Lonnqvist, J., Standertskjold-Nordenstam, C. G., Kaprio, J., Khaledy, M., Dail, R., Zoumalan, C. I., and Toga, A. W. (2001). Genetic influences on brain structure. *Nature Neuroscience*, 4:1253–1258.

Wainwright, M. J., Ravikumar, P., and Lafferty, J. D. (2007). High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In Schölkopf, B., Platt, J., and Hoffman, T., editors, *Advances in Neural Information Processing Systems 19*, pages 1465–1472. MIT Press.

Waldorp, L. (2009). Robust and unbiased variance of GLM coefficients for misspecified autocorrelation and hemodynamic response models in fMRI. *International Journal of Biomedical Imaging Volume 2009*, pages 1–11.

Weeda, W., Waldorp, L., Christoffels, I., and Huizenga, H. (2010). Activated region fitting: a robust high-power method for fMRI analysis using parameterized regions of activation. *Human Brain Mapping*, 30(8):2595–2605.

Wink, A. and Roerdink, J. (2006). BOLD noise assumptions in fMRI. *International Journal of Biomedical Imaging*, 2006:1–11.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Woodward, N. D., Rogers, B., and Heckers, S. (2011). Functional resting-state networks are differentially affected in schizophrenia. *Schizophrenia Research*, 130:86–93.

Worsley, K. (2001). Statistical analysis of activation images. In Jezzard, P., Matthews, P., and Smith, S., editors, *Functional MRI: An introduction to methods*, pages 251–270. Oxford University Press.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Zhang, X. and Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *The Annals of Statistics*, 39(1):174–200.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, 13:1059–1062.

Zhou, S., Lafferty, J., and Wasserman, L. (2010). Time varying undirected graphs. *Journal of Machine Learning Research*, 80(2–3):295–319.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36:1509–1533.